UNIVERSITY OF CALIFORNIA

SANTA CRUZ

SPHERICAL LATENT FACTOR MODEL FOR BINARY AND ORDINAL DATA

A dissertation submitted in partial satisfaction of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Xingchen Yu

December 2020

The Dissertation of Xingchen Yu
is approved:

Abel Rodriguez, Chair

Athanasios Kottas

Rajarshi Guhaniyogi

Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Spherical Latent Factor Model for Binary and Ordinal Data

by

Xingchen Yu

Factor models are widely used across diverse areas of application for purposes that include dimensionality reduction, covariance estimation, and feature engineering. Traditional factor models can be seen as an instance of linear embedding methods that project multivariate observations onto a lower dimensional Euclidean latent space. This thesis discusses a new class of geometric embedding models for multivariate binary and ordinal data in which the embedding space correspond to a spherical manifold, with potentially unknown dimension. The resulting models include traditional factor models as a special case, but provide additional flexibility. Furthermore, unlike other techniques for geometric embedding, the models are easy to interpret, and the uncertainty associated with the latent features can be properly quantified. These advantages are illustrated using both simulation studies and real data on voting records from the U.S. Congress as well as survey applications.

For my grandparents who rest in Heaven.

For my parents and my wife who always knew.

Acknowledgments

I would like to express my sincere gratitude towards my adviser Prof. Abel Rodriguez for guiding me through my Ph.D journey with his patience, vision and immense knowledge. I also would like to thank my thesis committee member, Prof. Athanasios Kottas and Prof. Rajarshi Guhaniyogi for their constructive and insightful critiques. I thank my friend and mentor Ernest Fokoue for his continuous encouragement and the door he kindly opened for me. I thank Dr.Yuanran Zhu for the discussion regarding the constrained derivative. I thank my fellow Ph.D colleagues Bohan Liu, Wenjie Zhao and Arthur Lui for their help and accompany.

Chapter 1

Introduction and Backgrounds

Factor analysis [1, 2] has a long history that dates back at least to the pioneering work of Charles Spearman on intelligence during the first decade of the 20^{th} century. Factor analysis is widely used across all sorts of disciplines within the social and natural sciences to account for and explain the correlations observed across multivariate responses. In traditional factor analysis, which in the case of normally distributed data includes principal component analysis as a special case, (the mean of) each response variable is represented as a linear combination of a common set of subject-specific *factors*. These factors can be interpreted as providing a linear embedding of the original high-dimensional data into a low-dimensional Euclidean space. These embeddings can be useful for various purposes, including data description, sparse estimation of covariance matrices, and/or as inputs to predictive models.

Linear embeddings on Euclidean spaces are relatively easy to compute and interpret. However, they might not be appropriate in all circumstances, particularly when the underlying data-generation mechanism involves highly non-linear phe-

nomena. Over the last 30 years, a rich literature has developed covering the use of nonlinear embeddings and data-driven, non-Euclidean embedding manifolds. Auto-encoders (e.g., see 3, 4, 5 and 6) are a natural generalization of principal component analysis. They explain the structure in the data through the composition of two non-linear functions (often represented as neural networks). The first function (usually called the *encoder*) embeds the input data into the low-dimensional latent variables, while the second function (called the *decoder*) maps those latent variables into the original space of the data in a way that minimizes the reconstruction error. Alternatively, locally-linear embeddings [7] reconstruct each input as a weighted average of its neighbours, and subsequently map the data into a low-dimensional embedding using those same weights. Isomap [8] extends the classical multidimensional scaling method into a general, data-driven manifold using geodesic distances measured on a neighbourhood graph. The goal is to preserve the intrinsic geometry of the manifold in which the data lies. Laplacian eigenmaps [9] also seek to preserve the intrinsic geometry of local neighborhoods. The embedding eigenvectors are obtained as the solution of a generalized eigenvalue problem based on the Laplacian matrix of the neighbourhood graph. Local tangent space alignment [10] finds the local geometry through the tangent space in the neighbourhood of inputs from which a global coordinate system for a general manifold is constructed. The inputs are embedded into a low-dimensional space through such global coordinate system. Finally, Gaussian process latent variable models [11] use Gaussian process priors to model the embedding function. These various approaches are very flexible in capturing the shape of the relationship between latent variables and outcomes, as well as the geometry of the underlying manifold on which the data lives. As a consequence, they can produce very compact representations of high-dimensional data using a very small number of

latent dimensions. However, interpreting and quantifying the uncertainty associated with these embeddings can be quite difficult because of identifiability issues. Indeed, when performing non-linear embeddings, invariance to affine transformations is not enough to ensure identifiability of the latent features. When the goal is prediction or sparse estimation of covariance matrices, this does not matter as these quantities are (usually) identifiable functions of the unidentified latent factors. However, when interest lies in the embedding coordinates themselves (as is the case in many social sciences applications), the lack of identifiability means that we are in the presence of irregular problems in which standard techniques for generating confidence/credible intervals are not applicable. Similarly, interpretation of the latent positions is dependent on the exact geometry of the embedding space, which is in turn only partially specified unless the lack of identifiability is addressed.

An alternative to the nonparametric embedding methods discussed above is to consider more flexible (but fixed) embedding spaces for which identifiability constraints can be easily derived. Such approach trades off some of the flexibility of the nonparametric methods for interpretability and the ability to properly quantify the uncertainty associated with the embedding. Along these lines, this thesis proposes a general framework for embedding multivariate binary and ordinal data into a general Riemannian manifold by exploiting the random utility formulation underlying binary regression models [12, 13]. To focus ideas, and because of their practical appeal, we place a special emphasis on spherical latent spaces.

Our focus on spherical latent spaces is driven by the needs of applications in a number of substantive areas. In political science, circular voting spaces have both theoretical and empirical support. In marketing applications, a spherical geometries can be used to explain the apparent lack of transitivity sometimes

observed in consumer behavior.

It is worthwhile noting that the literature has considered generalizations of data-reduction techniques such as principal components and factor analysis to situations in which the observations live on a manifold. Examples include principal Geodesic Analysis [14, 15, 16] and principal nested spheres [17]. This literature, however, is only marginally relevant to us since it is the parameters of our model, and not the data itself, that are assumed to live on a spherical manifold.

Next we review the traditional spatial voting models in the political science literature for which we also provide a brief background. In addition, we discuss recent "extremes voting together" phenomenon exhibits in this paradigm which motivates our adoption of the spherical manifold as the embedding space.

1.1 Motivation from the Political Science Literature

Spatial voting models [18, 19, 20, 21, 22, 23, 24] are widely used to estimate the preferences of legislators from roll call voting records, and have become an invaluable tool in the study of legislatures and other deliberative bodies. Spatial voting models aim to scale binary and polychotomous responses into a continuous (potentially multidimensional) linear scale, and are intimately related to traditional statistical tools for dimensionality reduction such as principal components and factor analysis. In the context of voting data, the latent space on which the responses are scaled is referred to as the *policy space*, while the latent traits are referred to as the *ideal points* of the legislators.

In one dimensional policy spaces, the ideal points generated by spatial voting models are often interpreted as capturing the ideology of the legislator on a liberal-conservative scale (e.g., see 20 and 25), with the ranking of the legislators in this scale typically becoming a key metric of interest. However, this interpretation can be suspect when the ideal points are learned exclusively on the basis of roll call votes. To address this issue, [26] used Early Day Motions (EDMs) instead of roll call records in the British House of Commons to learn about their ideology. EDMs are rarely debated and the Speaker of the House of Commons and Deputy Speakers generally do not sign EDMs. Another approach is to combine roll call data with other kinds of metadata. For example, [27], [28], [29], [30] and [31] develop methods that combine text and voting data to infer the ideology of legislators. In a similar spirit, [32] develop a method that uses manually-curated vote groups (such as those coming from the Policies Agenda Project, see 33) as metadata to infer issue-specific preferences.

Traditional spatial models rely on latent spaces endowed with Euclidean geometries, and therefore tend to work best in political systems in which the parties are relatively unified. One of the key motivations for the work on this dissertation is the analysis of voting records for legislatures in two-party systems in which parties are "fractious". In this kind of setting, it is common to see circumstances in which legislators that most observers would consider to be at opposite ends of the ideological spectrum vote together. [34] and [35] consider one example, namely, the first Blair government (1997-2001) in the United Kingdom. This government represented an uneasy alliance between a leadership that "had actively abandoned the tenants of socialist policy making and that had received a landslide mandate to rule" and a "historically and openly recalcitrant tranche of 'Old' Labour legislators, dismissive of the modernizing project in its entirety" [35]. In

the United States, the "Tea Conservatie Revolt" led by the Tea Party movement during the 2010 election [36, 37, 38], and the recent rise of the Justice Democrats during the 2018 election [39, 40] represent two more examples. Traditional spatial models fail in this setting where the "extremes vote together" because, under the Euclidean geometry, the "rebels" who sometimes vote with the opposition must necessarily be placed somewhere in the middle of the scale (e.g., see 34 and 35). Neither increasing the dimensionality of the latent space nor performing linear transformations of the latent space can address this issue (see Section 2.1).

In order to gain insights into legislatures in which the extremes vote together, [35] proposed a Bayesian non-parametric mixture model that identifies voting blocks within the U.K. House of Common. In a similar spirit, [41] and [42] developed random partition models for studying the voting record of the U.S. Supreme Court. While this kind of clustering models can provide valuable insights into the functioning of a deliberative body, they do not yield the kind of fine-grained ranking that has made spatial voting models so useful in practice. There is also an interesting literature focusing on the effect of the underlying utility functions on the estimates of the ideal points in Euclidean settings. For example, [43] describe a model in which the form of the utility function (quadratic or Gaussian on the Euclidean distance between points) is estimated from the data and conclude that extreme legislators are generally more sensitive to policy change than their more centrally located counterparts, while [44] discuss the use of the "city-block" (i.e., L^1) instead of L^2 distances in multidimensional spatial models, and [45] considers the general case of Minkowski (i.e., L^q) distances, where q is a parameter that is to be estimated from the data. More recently, [46] develop a model with non-monotone utility functions to explain the phenomenon of extremes voting together against the center. As an alternative, this thesis proposes a novel spatial voting

model that relies on a spherical policy space, and develops Bayesian inference procedures for it.

The idea that spherical policy spaces might be appropriate for representing political preferences dates back at least to [47], who provides a number of examples and notes that "circular shapes may be expected for alliance structures and for vote coalitions where extremists of the left and right coalesce for particular purposes". Our model can also be understood as operationalizing the so-called "horseshoe theory" (48; 49, pg. 118), which asserts that the far-left and the far-right are closer to each other than they are to the political center, in an analogous way to how the opposite ends of a horseshoe are close to each other. More generally, the use of spherical latent spaces for modeling preferences goes back at least to [50] and [51] in the economic and psychology literatures.

The bulk of the illustrations discussed in this manuscript will focus on the analysis of roll call votes in the U.S. Congress. Roll call votes are those in which each legislators votes "yea" or "nay" as their names are called by the clerk, so that the names of legislators voting on each side are recorded. It is important to note that not all votes fall into this category. Voice votes (in which the position of individual legislator are not recorded) are fairly common. This means that roll call votes represent a biased sample from which to infer legislator's preferences. Nonetheless, roll call data is widely used in political science. The U.S. Congress is composed of two chambers: the Senate and the House of Representatives. The U.S. Senate is composed of two representatives from each of the 50 States in the Union. On the other hand, the U.S. House of Representative consists of 435 voting members. Unlike the U.S. Senate, the delegation size of each state is proportional to its corresponding population. The total number of legislators included in each raw dataset might be slightly higher because of vacancies, which

are typically filled as they arise. While turnover in membership is typically quite low, these changes lead to voting records with (ignorable) missing values on votes that came up during the period in which a Senator was not part of the chamber. Additionally, missing values can also occur because of temporary absences from the chamber, or because of explicit or implicit abstentions. In the scope of this thesis, missing values were treated as if missing completely at random. While this assumption is not completely accurate (e.g., see 52), it is commonly made in most applied settings and we do not expect it to dramatically affect our analyses. Roll call data for the U.S. Congress can be obtained from `https://voteview.com/`.

1.2 Bayesian factor analysis models for binary data: A brief overview

Consider data consisting of independent multivariate binary observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$ associated with i *subjects*, where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,J})^T$ is a vector in which each entry is associated with a different *item*. In the political science application we discuss in Section 2.4 and 3.3, $y_{i,j}$ represents the vote of legislator (subject) i in question (item) j (with $y_{i,j} = 1$ corresponding to an affirmative vote, and $y_{i,j} = 0$ corresponding to a negative one). On the other hand, in marketing applications, $y_{i,j}$ might represent whether consumer (subject) i bought product (item) j (if $y_{i,j} = 1$) or not (if $y_{i,j} = 0$).

A common approach to modeling this type of multivariate data relies on a generalized bilinear model of the form

$$\Pr\left(y_{i,j} = 1 \mid \mu_j, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_i\right) = G\left(\mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\beta}_i\right), \tag{1.1}$$

where G is the (known) link function, and the intercepts $\mu_1, \ldots, \mu_J$ as well as the bilinear terms $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J \in \mathbb{R}^K$ and $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_I \in \mathbb{R}^K$ are all unknown and need to be estimated from the data (e.g., see 53, 54, 55 and 22). Different choices for the link function G lead to well-known classes models, such as logit and probit models. Furthermore, given a distribution over the latent factors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_I$, integrating over them (which, in the case of binary data, can be done in closed form only in very special cases) yields a wide class of correlation structures across items.

The class of factor analysis models for binary data in (1.1) can be derived through the use of random utility functions [12, 13]. To develop such derivation, we interpret the latent trait $\boldsymbol{\beta}_i \in \mathbb{R}^K$ as representing the preferences of subject i over a set of unobserved item characteristics, and associate with each item two positions, $\boldsymbol{\psi}_j \in \mathbb{R}^K$ (corresponding to a positive responsive, i.e., $y_{i,j} = 1$) and $\boldsymbol{\zeta}_j \in \mathbb{R}^K$ (corresponding to a negative one, i.e., $y_{i,j} = 0$). Assuming that individuals make their choice for each item based on the relative value of two random quadratic utilities,

$$U_+(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i) = -\left\|\boldsymbol{\psi}_j - \boldsymbol{\beta}_i\right\|^2 + \epsilon_{i,j}, \qquad U_-(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i) = -\left\|\boldsymbol{\zeta}_j - \boldsymbol{\beta}_i\right\|^2 + \nu_{i,j}, \tag{1.2}$$

where $\epsilon_{i,j}$ and $\nu_{i,j}$ represent random shocks to the utilities, and $\upsilon_{i,j} = \nu_{i,j} - \epsilon_{i,j}$ are independently distributed for all i and j and have cumulative distribution function $G_j(x) = G(x/\sigma_j)$, it is easy to see that

$$\Pr\left(y_{i,j} = 1 \mid \boldsymbol{\psi}_j, \boldsymbol{\zeta}_j, \boldsymbol{\beta}_i, \sigma_j\right) = \Pr\left(U_+\left(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i\right) > U_-\left(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)\right) = G\left(\mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\beta}_i\right),$$

where $\boldsymbol{\alpha}_j = 2(\boldsymbol{\psi}_j - \boldsymbol{\zeta}_j)/\sigma_j$ and $\mu_j = (\boldsymbol{\zeta}_j^T \boldsymbol{\zeta}_j - \boldsymbol{\psi}_j^T \boldsymbol{\psi}_j)/\sigma_j$. This formulation, which is tightly linked to latent variable representations used to fit categorical models (e.g., see 56), makes it clear how the factor model embeds the binary observation into a Euclidean latent space. Furthermore, this construction also highlights a number

of identifiability issues associated with the model. In particular, note that the utility functions in (4.14) are invariant to affine transformations. Therefore, the positions $\boldsymbol{\psi}_j$, $\boldsymbol{\zeta}_j$ and $\boldsymbol{\beta}_i$ are only identifiable up to translations, rotations and rescalings, and the scaling σ_j cannot be identified separately from the scale of the latent space. These identifiability issues are usually dealt with by fixing $\sigma_j = 1$ for all $j = 1, \ldots, J$, and by either fixing the position of $K+1$ legislators (e.g., see 23), or by fixing the location and scale of the ideal points, along with constraints on the matrix of discrimination parameters (e.g., see 57). In either case, identifiability constraints can be enforced either a priori, or a posteriori using a parameter expansion approach (e.g., see 58, 59 and 60).

Bayesian inference for this class of models requires the specification of prior distributions for the various unknown parameters. For computational simplicity, it is common to use (hierarchical) priors for the μ_js, $\boldsymbol{\alpha}_j$s and $\boldsymbol{\beta}_i$s, so that computation can proceed using well-established Markov chain Monte Carlo algorithms (e.g., see 56 and 61). However, the hyperparameters of these priors need to be selected carefully. For example, when the dimension K of the embedding space is unknown and needs to be estimated from the data, it is important to ensure that the prior variance on $\theta_{i,j} = G(\mu_j + \boldsymbol{\alpha}_j^T \boldsymbol{\beta}_i)$ induced by these priors remains bounded as $K \to \infty$ if one is to avoid Bartlett's paradox [62]. One approach to accomplish this goal is to select the prior covariance matrix for the $\boldsymbol{\beta}_i$s to be diagonal, and to have the marginal variance of its components decrease fast enough as more of them are added.[1] This type of construction, which is related but distinct from the one implied by the Indian Buffet process [63], provides a prior that is consistent as the number of components K grows, and which can be interpreted

[1] As an example, consider a probit model (i.e., $G(x) = \Phi(x)$) where, a priori, $\mu_j \sim N(0, 1/2)$, $\alpha_{j,k} \sim N(0, 1/2)$, and $\beta_{i,k} \sim N(0, 6/[\pi k]^2)$. In this case, $\theta_{i,j} = \Phi\left(\mu_j + \sum_{k=1}^{K} \alpha_{j,k}\beta_{i,k}\right)$ converges in distribution to a uniform distribution on $[0, 1]$ as $K \to \infty$ (see Appendix A.2.

as a truncation of an infinite dimensional model. Assigning prior distributions to the $\boldsymbol{\psi}_j$s and $\boldsymbol{\zeta}_j$s (which, in turn, imply priors on the $\boldsymbol{\alpha}_j$s and μ_js) is also a possibility, but it is rarely done in practice. In that setting, ensuring a satisfactory behavior as the number of dimensions grows typically requires that the variances of all three latent positions decrease as the number of dimensions of the latent space increases. This observation will be important to some of the developments in Section 3.1.2.

1.3 Posterior inference techniques on Riemannian Manifolds

The posterior distributions associated with the models proposed in this thesis are analytically intractable and necessitate the use of approximate technique such as Markov chain Monte Carlo (MCMC) algorithms. Furthermore, key parameters of interest live in non-Euclidean spaces, which necessitates that we move beyond traditional random walk Metropolis-Hastings algorithms. This section reviews Hamiltonian Monte Carlo (HMC) algorithms [64] , with a particular emphasis on versions of the algorithm that are applicable in general Riemannian manifolds. These types of algorithms will be critical for posterior inference in the classes of spherical factor models we discuss in this dissertation.

1.3.1 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) [64], also known as hybrid Monte Carlo, is a gradient-based sampling method that is widely used in both statistical and

computer science literature. It could be considered as a Metropolis-Hastings algorithms that uses the Hamiltonian dynamics to generate proposals rather than using a random work. Thanks to the Hamiltonian dynamics, the samples generated are less correlated than the traditional Metropolis-Hastings. When posterior full conditional distributions do not belong to a known family, or the choice of conjugate prior to force posterior conjugacy is not justified, HMC serves as a great method of generating posterior samples. After assuming a predetermined leapfrog steps L, step size ϵ and a momentum distribution $q(\phi)$, the HMC algorithm proceeds sequentially for each iteration t as follows,

1. Sample ϕ from the momentum distribution $q(\phi)$.
2. Make a half step for ϕ and a full step for θ:
   (a) $\phi \leftarrow \boldsymbol{\phi} + \frac{1}{2}\epsilon \frac{\partial \log\, p(\theta|y)}{\partial\, \theta}$.
   (b) $\theta \leftarrow \theta + \epsilon \frac{\partial \log\, q(\phi)}{\partial\, \phi}$.
3. For each of the following $L-1$ leap frog steps:
   (a) $\phi \leftarrow \phi + \epsilon \frac{\partial \log\, p(\theta|y)}{\partial\, \theta}$.
   (b) $\theta \leftarrow \theta + \epsilon \frac{\partial \log\, q(\phi)}{\partial\, \phi}$.
4. Make the last half step for ϕ to complete the leapfrog steps:
   (a) $\phi \leftarrow \phi + \frac{1}{2}\epsilon \frac{\partial \log\, p(\theta|y)}{\partial\, \theta}$.
5. Accept and reject the resulting θ based on the following step.

(a) Compute

$$R \leftarrow \frac{p(\theta \mid y)q(\phi)}{p(\theta^{t-1} \mid y)q(-\phi^{t-1})}. \tag{1.3}$$

(b) set

$$\theta^t = \begin{cases} \theta & \text{with probability of } \min(R,1) \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

One key challenge associated with the use of HMC algorithms is the need to tune the step size ϵ and the number of leaps L, as well as the hyperparameter for the momentum distribution.

Practically speaking, the choice of of leapfrog steps L and step size ϵ and their interaction greatly affect the acceptance ratio and the correlation between posterior samples. Theory [65] suggests that HMC is optimally efficient when the acceptance ratio is roughly 65%. This should be contrasted against a 23% optimal acceptance ratio [66] for multidimensional Metropolis-Hastings algorithms. Further analysis [67] employed the underlying geometry of HMC to construct an automatic selection criterion of the step size, which leads to immediate results showing the target acceptance ratio of 65% could be relaxed to approximately 60% to 90% with larger values more robust in practical applications.

With these knowledge in mind, leapfrog steps L and step size ϵ should be ideally tuned by targeting the acceptance ratio to be between 60% and 90% and then use performance measures such as effective sample size, posterior log-likelihood to compare the performance of various combination of leapfrog steps L and step

size ϵ. However, in addition to L and ϵ, the hyper parameter of the momentum distribution needs also be tuned. By convention, often the Gaussian distribution is chosen as the momentum distribution which leads to the need to tune the covariance matrix. Tuning the covariance matrix could be a very challenging task especially for high dimensional parameter space and thus diagonal covariance matrix is often selected to bypass this problem. However, this simplification will result in non-optimal Hamiltonian transition towards the target distribution.

As such, various methods have been proposed to tackle this problem. Riemannian manifolds HMC (RM-HMC) [68] takes into account the local Riemannian structure of the parameter space by using the Fisher information matrix as the precision matrix for $q(\phi)$. This eliminates the need to tune this parameter and improves the mixing of the algorithm. Geodesic HMC [69] (GHMC) improves from RM-HMC by using the splitting of Hamiltonian techniques and embedding representation to avoid the troublesome numeric integrator. Adaptive HMC [70] uses an adaptive method such that hyper parameters could be automatically tuned based on target acceptance ratio. Relativistic HMC [71] introduces a novel method of using the speed of light as the maximum speed it could travel to the target distribution and also provides a correspondence between this extension of HMC and Stochastic gradient descent using momentum. [72] provides an interesting observation of the relation between slice sampling and HMC. In the scope of this document, we will focus our discussion on the RM-HMC and its extension GHMC since the key parameters of our proposed models lives on a Riemannian manifold.

1.3.2 Riemannian Manifold HMC

Riemannian manifold HMC (RM-HMC) [68] was proposed to tackle high dimensional, strongly correlated posterior distributions. It uses the Fisher Information associated with the target distribution as the precision matrix of the Gaussian distribution of the momentum variables. This provides a fully automated adaptation process that allows exploration of different likelihood region in the Riemannian structure of the parameter space, which gives HMC a deterministic mechanism of HMC to auto-tune the covariance matrix of the proposed distribution. The algorithm uses a semi-explicit second order symplectic integrator for non-separable Hamiltonian equations.

The parameter space of a statistical model exhibits a Riemannian structure whose invariant metric tensor $\boldsymbol{G}(\boldsymbol{\theta})$ is defined by the non-degenerate Fisher Information $E\{\nabla_\theta \mathcal{L}(\boldsymbol{\theta} \mid Y)\nabla_\theta \mathcal{L}(\boldsymbol{\theta} \mid Y)\}^T$ where $\mathcal{L}(\boldsymbol{\theta} \mid Y) = \log p(\boldsymbol{\theta} \mid Y)$ [73] . Under the conventional Gaussian momentum distribution, the Hamiltonian dynamics defined on the Riemannian manifold incorporating the metric tensor become the following:

$$H(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta} \mid Y) + \frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{\phi}, \tag{1.4}$$

where $p(\boldsymbol{\theta} \mid Y) = -\mathcal{L}(\boldsymbol{\theta} \mid Y) + \frac{1}{2}\log(2\pi)^D|\boldsymbol{G}(\boldsymbol{\theta})|$. Consequently the Hamiltonian dynamics with the matrix tensor becomes:

$$\frac{\partial H}{\partial \phi_i} = (\boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{\phi})_i, \tag{1.5}$$

$$-\frac{\partial H}{\partial \theta_i} = \frac{\partial \mathcal{L}(\boldsymbol{\theta} \mid Y)}{\partial \theta_i} - \frac{1}{2}Tr\left[\boldsymbol{G}(\boldsymbol{\theta})^{-1}\frac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \theta_i}\right] + \frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{G}(\boldsymbol{\theta})^{-1}\frac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \theta_i}\boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{\phi}. \tag{1.6}$$

Evidently, unlike the vanilla HMC, the Hamiltonian dynamics now defined on the Riemannian manifold are no longer factorisable into two equations and there is now a coupling between the momentum $\boldsymbol{\phi}$ and kinetic variable $\boldsymbol{\theta}$ due to the incorporation of the metric tensor $\boldsymbol{G}(\boldsymbol{\theta})$. Therefore, a numerical symplectic numerical integrator for solving this non-separable Hamiltonian is required to ensure the posterior converges to the correct distribution. The detail of this numerical symplectic integrator is shown in the appendix of [68]. The main disadvantage of RM-HMC algorithms is that this simplectic integrator is computationally inefficient, requiring $\mathcal{O}(N^3)$ operations, where N is the sample size. This feature make the algorithm impractical to apply to large datasets with high dimensional parameter spaces.

A thorough and detailed analysis of RM-HMC comparing with various algorithms on 5 datasets could be found in [68]. The largest number of dimension studied is merely 24 which is much smaller than our smallest dataset that contains 600 parameters. Due to the concern of the scalability, we determined RM-HMC is not suitable in our spherical latent factor model. Fortunately, GHMC improves the RM-HMC algorithm by using the splitting method on the Hamiltonian dynamics and with the embedded transformation, it is able to bypass the need to use the troublesome numeric symplectic integrator altogether.

1.3.3 Geodesic Hamiltonian Monte Carlo

An additional shortcoming of RM-HMC is that it requires a global coordinate system, which is not available for manifolds such as great spheres in which artificial boundaries should be induced. Sampling from manifolds is generally considered a challenging problem that has not receive much attention. A recent introduc-

tion to geometric measure theory by [74] illustrated the challenges associated with sampling from the manifolds. Geodesic Hamiltonian Monte Carlo (GHMC) [69] provides a straightforward tool for sampling from distributions defined on manifolds. In contrast to the RM-HMC, it avoids a complex numeric integrator and does not require a global coordinate system. It also provides a scalable and efficient way to obtain samples from target distributions defined on manifolds that can be embedded in Euclidean space, by exploiting the existing forms of the geodesics such as spheres, affine subspaces, Stiefel manifolds or product manifolds.

The Hausdorff measure is a fundamental concept in geometric measure theory, and can be mathematically defined in terms of a limit of coverings of the manifold [74]. Heuristically, for a manifold embedded in $\mathbb{R}^n$, it can be considered as the surface area of the manifold. The relationship between the m-dimensional Hausdorff measure $\mathcal{H}^m$ and the Lebesgue measure λ^m on $\mathbb{R}^M$ is defined by the area formula [75], which can be naturally extended to Riemannian manifolds, where

$$\mathcal{H}^m(d\boldsymbol{\theta}) = \sqrt{|\boldsymbol{G}(\boldsymbol{\theta})|}\lambda^m(d\boldsymbol{\theta}). \tag{1.7}$$

This extension should be familiar to Bayesians as Jeffreys prior, where $\boldsymbol{G}$ again represents the Fisher information. Hence the Hausdorff measure becomes a natural reference measure that allows reparameterization without the need to compute the determinant of Jacobian matrix.

Writing the log target density with respect to the Hausdorff measure of the manifold in 1.4, we obtain

$$H(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\log \pi_{\mathcal{H}}(\boldsymbol{\theta} \mid Y) + \frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{\phi}, \tag{1.8}$$

where $\log \pi_H(\boldsymbol{\theta} \mid Y)$ is the log density of the posterior distribution with respect to the Hausdorff measure. To solve the Hamiltonian dynamics in (1.8), we can use the splitting method shown in [64]. The first component of this splitting is the potential energy term $H^{[1]}(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\log \pi_H(\boldsymbol{\theta} \mid Y)$ which has exact solution of a linear update of the momentum $\boldsymbol{\phi}$. The second splitting component is just the kinetic energy term $H^{[2]}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{\phi}$ which is simply a Hamiltonian without any potential energy. The solution of this Hamiltonian equation can be easily shown to be a geodesic flow under the Levi-Civita connection of $\boldsymbol{G}$ [76]. In addition, we can represent this algorithm in terms of its embedding, which altogether bypasses the need to explicitly compute the computationally troublesome metric tensor $\boldsymbol{G}$ and the need to have a global coordinate system [69]. Hence after assuming a predetermined leapfrog steps L and step size ϵ, the corresponding GHMC algorithm also proceeds sequentially for each iteration t as follows,

1. Sample $\boldsymbol{\phi} \sim N(0, \boldsymbol{I}_d)$.
2. Set $\boldsymbol{\phi} \leftarrow \Big(\boldsymbol{I}_d - N(\boldsymbol{\theta})N(\boldsymbol{\theta})^T\Big)\boldsymbol{\phi}$.
3. For each of the L leap steps:
   (a) $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla \log \pi_{\mathcal{H}}(\boldsymbol{\theta} \mid y)$.
   (b) $\boldsymbol{\phi} \leftarrow \Big(\boldsymbol{I}_d - N(\boldsymbol{\theta})N(\boldsymbol{\theta})^T\Big)\boldsymbol{\phi}$.
   (c) Update $(\boldsymbol{\theta}, \boldsymbol{\phi})$ by its geodesic flow for a step size of ϵ.
   (d) $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla \log \pi_{\mathcal{H}}(\boldsymbol{\theta} \mid y)$.
   (e) $\boldsymbol{\phi} \leftarrow \Big(\boldsymbol{I}_d - N(\boldsymbol{\theta})N(\boldsymbol{\theta})^T\Big)\phi$.

4. Compute

$$\begin{aligned} h \leftarrow & log\, \pi_{\mathcal{H}}(\boldsymbol{\theta} \mid y) - \frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\phi}, \\ h^{t-1} \leftarrow & \log \pi_{\mathcal{H}}(\boldsymbol{\theta}_{t-1} \mid y) - \frac{1}{2}\boldsymbol{\phi}_{t-1}^T\boldsymbol{\phi}_{t-1}, \\ R \leftarrow & \exp\left(h - h^{t-1}\right). \end{aligned}$$

5. Set

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta} & \text{with probability of } \min(R, 1), \\ \boldsymbol{\theta}^{t-1} & \text{otherwise,} \end{cases}$$

where $\boldsymbol{I}_d$ is the identity covariance matrix of size d and $\boldsymbol{I}_d - N(\boldsymbol{\theta})N(\boldsymbol{\theta})^T$ is the orthogonal projection to the tangent space of $\boldsymbol{\theta}$. In the scope of this document, the embedded manifolds of interest are spheres, and the normal to the tangent space $N(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\left(\boldsymbol{I}_d - N(\boldsymbol{\theta})N(\boldsymbol{\theta})^T\right)\phi$ is an orthogonal projection of an arbitrary $\boldsymbol{\phi}$ onto the tangent space. Since the geodesics of the spheres are the great circle rotations with respect to the origin, the corresponding geodesic flows are defined as,

$$\begin{aligned} \boldsymbol{\theta}(t) =& \boldsymbol{\theta}(0)\cos(\alpha t) + \frac{\boldsymbol{\phi}(0)}{\alpha}\sin(\alpha t), \\ \boldsymbol{\phi}(t) =& \boldsymbol{\phi}(0)\cos(\alpha t) - \alpha\,\boldsymbol{\phi}(0)\sin(\alpha t), \end{aligned} \tag{1.9}$$

where $\alpha = \|\boldsymbol{\phi}(0)\|$ is the constant angular velocity at time 0. Other than the evaluation of gradient of the log-density and its associated gradient, the algorithm requires only vector-vector multiplications and additions, hence it is rather efficient. In other words, this algorithms scales linearly in d. In addition, because of the way the embedding is constructed, $\boldsymbol{\phi}$ can be sampled from a multivari-

ate Gaussian distribution with an Identity covariance matrix of size d to allow orthogonal idempotent projection. As a result, the need to tune the covariance matrix has been completely eliminated, which is another key advantage of this algorithm. Therefore, we employ GHMC to carry out posterior inference for our models developed in this thesis.

1.4 Thesis structure

The first part of the thesis up so far lay the ground work to introduce our Bayesian spherical latent factor model for binary data. In Chapter 2, we discuss the univariate dimensional version of our model which generalizes its Euclidean counterpart and we show proof-of-concept to justify the use of circular space with the roll-call voting data from the U.S Congress between 1988 to 2019. In Chapter 3, we build upon our univariate model to generate a general framework for embedding binary data into spherical latent spaces on which a new class of prior distributions that does not degenerate as dimension increases is proposed. In Chapter 4, we extend our proposed model to also embed ordinal data into the spherical latent space. This is achieved through a careful selection of different types of ordinal structure and a comparison of various configurations of the link function. In Chapter 5, we conclude the thesis and identify several areas for potential future works in this line of research.

All the implementation of the models developed in this thesis is available at `https://github.com/Xingchen-Yu`.

Chapter 2

Circular Factor Model for Binary Data

This chapter introduces a new class of spatial voting models in which preferences live in a circular space. Our formulation includes the one-dimensional version of the Euclidean model discussed in Section 1.2 as a special (limiting case), allowing the data to inform us about the geometry of the underlying latent space. As we discussed in Section 1.1, a circular structure for the latent space is motivated by both theoretical (the so-called "horseshoe theory" of political thinking) and empirical (goodness of fit) considerations in which members at the opposite ends of the ideological spectrum reveal similar preferences by voting together against the rest of the legislature. In particular, by applying the model to roll-call voting data from the U.S. Congress between 1988 and 2019, we demonstrate that circular latent spaces provide a better explanation for the political process in the House of Representatives than Euclidean models, that policy spaces have become increasingly circular in recent years (and, especially, since 2010), and that legis-

lators's rankings generated through the use of the circular geometry tend to be more consistent with their stated policy positions.

We start this chapter by a motivating example from the traditional roll-call voting record analysis of the 116$^{\text{th}}$ Congress. Next we introduces our circular factor model and discusses the link function, prior elicitation, identifiability and its connection to the traditional Euclidean factor model. Section 2.3 discusses our computational approach to estimating model parameters, which is based on Geodesic Hamiltonian Monte Carlo (GHMC) algorithms. Section 2.4 illustrates the behavior of our model on the roll call data of the 112 and 116$^{\text{th}}$ U.S. House of Representatives. In addition, we also present a longitudinal analysis of the contemporary U.S. House of Representatives. Finally, Section 2.5 summarizes the chapter.

2.1 A motivating example: Ranking "The Squad"

The November 2018 midterm election saw the Democratic Party win a new majority in the House of Representatives on the back of a record number of women, young, and minority candidates. Particularly notable among them is a group of four new members (Alexandria Ocasio-Cortez of New York, Ilhan Omar of Minnesota, Ayanna Pressley of Massachusetts, and Rashida Tlaib of Michigan, all women of color under 50 supported by the Justice Democrats political action committee), who often refer to themselves as "The Squad". As discussed in [39] and [40], the Squad is widely understood to belong to the left wing of the Democratic party, supporting policies such as the Green New Deal, reparations for slavery, and abolishing the Immigration and Customs Enforcement Agency. Partly because of their support for these policies, they have shown a willingness

to challenge the leadership of their party and to vote against it on some issues.

Table 2.1 presents the rank order of the members of the Squad on a liberal-conservative scale based on their voting record during the first session of the 116$^{\text{th}}$ Congress (extending between January 3, 2019 and January 3, 2020). These rankings were obtained by fitting one- and two-dimensional versions of the Euclidean model described in [22] and [23] (see also Section 1.2). Counterintuitively, all members of the Gang are ranked towards the center of the political spectrum under both models. Most importantly, note that the addition of a second dimension does not dramatically affect the original surprising conclusion that they all appear to belong to the conservative wing of the Democratic party. As we discussed in the introduction, this counterintuitive result is a direct consequence of the Euclidean geometry underlying these models: If a legislator votes with the opposite party against the majority of its own, the only possible explanation is that the legislator is a moderate.

Table 2.1: Median rank of the members of the "Squad" during the first session of the 116$^{\text{th}}$ U.S. House of Representatives according to two scaling models: A one dimensional Euclidean model, and a two dimensional Euclidean model. In the case of the two dimensional model, the ranking provided is along the first (highest variability) dimension of the policy space. Lower numbers for the ranks correspond to more liberal legislators. Numbers in parenthesis correspond to 95% credible intervals.

	Rank Order	
	Euclidean (1D)	Euclidean (2D)
Pressley (D MA-7)	172 (131,196)	168 (113,200)
Omar (D MN-5)	176 (135,198)	160 (98,195)
Tlaib (D MI-13)	180 (146,200)	169 (108,200)
Ocasio-Cortez (D NY-14)	203 (185,215)	197 (172,213)

To further investigate the voting behavior of the Squad, we also fitted the non-parametric mixture model described in [35] to these data. The model identifies three groups of Democrats that appear to have distinct behavior: a small group of 18 legislators representing some of the districts that were flipped by Democrats during the 2018 election and whose seats are widely understood to be at most risk in the 2020 election (we could call these the *vulnerables*), a medium sized group of 61 legislators that include most of the remaining representatives from flipped districts as well as a number of legislators with relatively short tenures in the House (we could call them the *pragmatists*), and a large group of 155 legislators that includes the leadership as well as most representatives with a long tenure in the House (call them the *establishment*). Interestingly for our purposes, the members of the Squad are not split off into a separate group that includes left-wing activists, but are instead clustered with the establishment. Note that, because of the structure of the [35] model, no further rankings of the legislators are possible within each block.

2.2 Bayesian spatial voting models with circular policy spaces

The framework described in Section 1.2 lends itself naturally to extensions to policy spaces with more general geometric properties. In particular, we can embed the latent positions $\boldsymbol{\beta}_i$, $\boldsymbol{\psi}_j$ and $\boldsymbol{\zeta}_j$ on a Riemannian manifold $\mathcal{D}$, and then replace the Euclidean distance used in the definition of the utility functions in Equation

(4.14) with the geodesic distance ρ on $\mathcal{D}$, so that

$$U_{\text{Yea}}(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i) = -\rho\left(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i\right)^2 + \epsilon_{i,j}, \quad U_{\text{Nay}}(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i) = -\rho\left(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)^2 + \nu_{i,j}. \quad (2.1)$$

In this chapter, we focus on the special case where $\mathcal{D}$ corresponds to the unit circle, so that $\beta_i, \psi_j, \zeta_j \in [-\pi, \pi]$ can be interpreted as angular positions on the circle, and $\rho(a, b) = \arccos(\cos(a - b))$ is just the smallest angle separating a and b. Because of the conditional independence among votes, this formulation leads to a likelihood function of the form

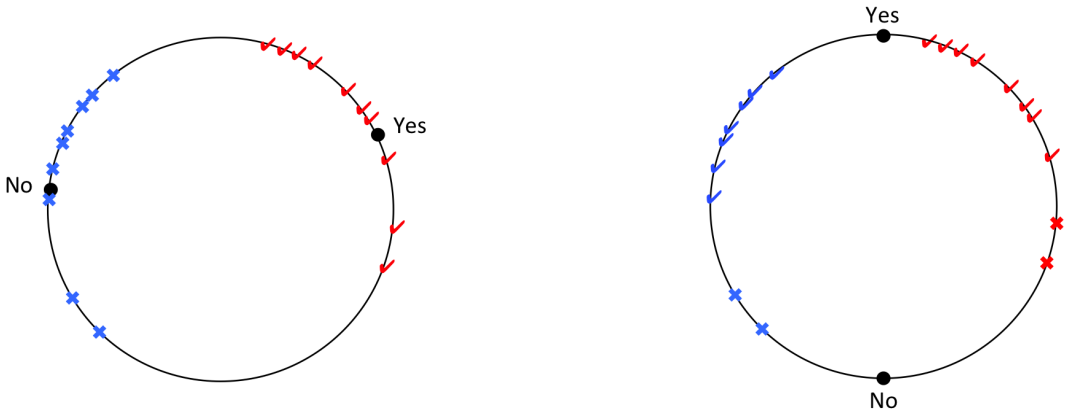
$$\Pr(\boldsymbol{Y} \mid \boldsymbol{\psi}, \boldsymbol{\zeta}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left[G_{\kappa_j}\left(e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right)\right)\right]^{y_{i,j}} \left[1 - G_{\kappa_j}\left(e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right)\right)\right]^{1-y_{i,j}}, \quad (2.2)$$

where $\boldsymbol{Y}$ is the $I \times J$ data matrix with entries $y_{i,j}$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_J)^T$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_J)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)^T$ are the vectors of unknown positions for all legislators and questions in the policy space, G_{κ_j} is the cumulative distribution function associated with $\nu_{i,j} - \epsilon_{i,j}$, and

$$e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right) = \{\arccos(\cos(\zeta_j - \beta_i))\}^2 - \{\arccos(\cos(\psi_j - \beta_i))\}^2.$$

Figure 2.1 provides some intuition for the additional flexibility intrinsic to the circular voting model, and in particular, for its ability to accommodate situations in which the "extremes voting together". The left panel depicts a situation in which the outcome of a vote follows along party lines. This type of situation, in which the "Yea" and "Nay" positions fall close to the center of mass of opposite parties, represents the most typical type of question in most legislatures, and is

well modeled using traditional Euclidean policy spaces. In particular, it captures situations in which moderates from one party vote with the other party. In contrast, the right panel depicts a situation in which moving the "Yea" and "Nay" positions to the upper and lower poles leads, with the same ideal points as before, to an outcome in which the "extreme" members of each party join forces in voting against the question.



(a) Partisan vote **(b)** Extremes votes together

Figure 2.1: Two configurations in a circular policy space. Check marks and crosses correspond to the ideal points of legislators voting in favor and against a question, and are the same on both panels. Circles correspond to the "Yea" and "Nay" positions for the questions. The left panel, in which the bill positions are located in the upper hemisphere, corresponds to a vote along party lines. The right panel, in which the "Yea" and "Nay" positions fall in the upper and lower poles, corresponds to a question in which the extremes vote together.

2.2.1 Link function

Selecting a link function for the model is non-trivial. Note that, unlike the Euclidean distance in $\mathbb{R}^q$, the geodesic distance on the circle takes values in the interval $[0, \pi]$. This means that the difference between two squared distances takes values in $[-\pi^2, \pi^2]$, and our link function must account for this. We propose to

define G_{κ_j} as the cumulative distribution function of scaled and shifted symmetric beta distribution with density,

$$g_{\kappa_j}(z) = \frac{1}{2\pi^2}\frac{\Gamma(2\kappa_j)}{\Gamma(\kappa_j)\Gamma(\kappa_j)}\left(\frac{\pi^2+z}{2\pi^2}\right)^{\kappa_j-1}\left(\frac{\pi^2-z}{2\pi^2}\right)^{\kappa_j-1}, \quad z\in[-\pi^2,\pi^2]. \tag{2.3}$$

The use of this transformed symmetric beta distribution has two advantages in this setting. First, the parameter κ_j has a direct interpretation as a precision parameter. Indeed, the variance of a random variable with density (4.26) is equal to $\pi^4/(2\kappa_j+1)$. This provides a direct analogy with the scaling parameter σ_j introduced in Section 1.2. Secondly, note that

$$\lim_{\kappa_j\to\infty}\frac{g_{\kappa_j}(z)}{\sqrt{\frac{2\kappa_j+1}{2\pi^5}}\exp\left\{-\frac{2\kappa_j+1}{\pi^4}z^2\right\}} = 1, \tag{2.4}$$

i.e., as the concentration parameter increases, the density g_{κ_j} resembles that of a normal distribution with zero mean and variance $\pi^4/(2\kappa_j+1)$. This limiting behavior will play an important role when discussing the relationship between our circular model and traditional Euclidean models (see Section 2.2.4).

2.2.2 Identifiability

As mentioned in the introduction, the goal of our circular model is scaling rather than prediction. Thus, the identifiability of the latent traits $\beta_1,\ldots,\beta_I$ is crucial. We discuss here the constraints required to make all model parameters identifiable.

To start, notice that the likelihood in Equation 2.2 remains constant if the same shift is applied to all β_is, ψ_js and ζ_js. We address this location invariance through a careful selection of the prior distribution on the β_is (see Section 2.2.3). Further-

more, the likelihood also remains constant if any angle is independently increased or decreased by 2π. This invariance to "wrappings around the circle" is easily addressed by mapping all angles to the $[-\pi, \pi]$ interval. Finally, note that the model is invariant to reflections of the policy space, just like the one-dimensional Euclidean model. We address this by fixing the sign of the ideal point of one particular legislator (e.g., the whip of one of the parties).

A key difference between the Euclidean and circular models, however, is that the positions in the circular model are not invariant to changes in scale. As a consequence, the parameters $\kappa_1, \ldots, \kappa_J$ controlling the variance of the link function in Equation (2.2) are identifiable and can be estimated separately from the β_is, ψ_js and ζ_js. In fact, because the geodesic distance ρ is bounded, learning κ_js from the data and allowing them to vary across questions is key to accommodate the full variety of voting behaviors, and in particular, unanimous votes.

2.2.3 Prior distributions

We consider now the selection of priors on the latent positions. Since β_i, ζ_j and ψ_j represent angles, it is natural to use independent von Mises distributions for these parameters,

$$\begin{aligned} \beta_i \mid \omega_\beta, \tau_\beta &\sim \text{vonMis}(\tau_\beta, \omega_\beta), \\ \psi_j \mid \omega_\psi, \tau_\psi &\sim \text{vonMis}(\tau_\psi, \omega_\psi), \\ \zeta_j \mid \omega_\zeta, \tau_\zeta &\sim \text{vonMis}(\tau_\zeta, \omega_\zeta). \end{aligned} \tag{2.5}$$

A random variable Z follows a von Mises distributions with mean τ and concen-

tration ω, $Z \sim \text{vonMis}(\tau, \omega)$, if it has density

$$p(z) = \frac{1}{2\pi I_o(\omega)} \exp\{\omega \cos(z - \tau)\}, \qquad z \in [-\pi, \pi],$$

where $I_k(\omega)$ is the modified Bessel function of order k. When $\omega = 0$, the von Mises distribution becomes the uniform distribution on the circle. On the other hand, as ω grows, the distribution behaves as a normal distribution with variance $1/\omega$, i.e.,

$$\lim_{\omega \to \infty} \frac{\frac{1}{2\pi I_o(\omega)} \exp\{\omega \cos(z - \tau)\}}{\sqrt{\frac{\omega}{2\pi}} \exp\left\{-\frac{\omega}{2}(z - \mu)^2\right\}} = 1. \tag{2.6}$$

In fact, we can think about the von Mises as being equivalent to the Gaussian distribution on the circle.

To elicit the hyperparameters of the model we rely on the intuition provided by Figure 2.1. Since we want the question's positions to potentially be located anywhere on the circle, we set $\omega_\psi = \omega_\zeta = 0$ (leading, as we mentioned before, to uniform prior on the circle for these two parameters). On the other hand, the ideal points are assigned a zero mean, i.e., $\tau_\beta = 0$, and a non-zero precision, i.e, $\omega_\beta > 0$. This structure ensures (weak) identifiability of all latent positions to location shifts (recall our discussion from Section 2.2.2). In particular, we let ω_β be an Exponential hyperprior with mean $\theta = 10$, so that $\Pr(-\pi/2 < \beta_i < \pi/2) \approx 0.95$, and perform a sensitivity analyses. This choice for ω_β reflects our prior belief that a Euclidean model is reasonable in most cases, and therefore most ideal points will be concentrated on the upper hemisphere (see discussion in Section 2.2.4). Finally, we assume that the κ_js are independent and identically distributed a priori from an exponential prior with mean λ, which is in turn assigned a (conditionally conjugate) inverse Gamma prior with one degree of freedom and rate parameter

$\xi = 25$ (i.e., $1/\lambda$ follows an exponential distribution with mean $\xi = 25$). Again, we investigate the impact of this choice in our applications through a sensitivity analysis.

2.2.4 Relationship with traditional Euclidean models

The probit version of the one-dimensional Euclidean model described in Section 1.2 can be seen as a special (limit) case of our circular model. To understand this relationship, consider projecting the latent angles that describe the circular model onto the tangent space at 0 (see Figure 2.2). Two points need to be made about such projection. First, note that as $\omega_\beta \to \infty$, the ideal points of the legislators will tend to concentrate around the point of tangency. As a consequence, the projection of the angles β_i, ψ_j and ζ_j onto the tangent space (labeled β_i', ψ_j' and ζ_j' in the figure) satisfy $\beta_i' = \tan\beta_i \approx \beta_i$, $\psi_j' = \tan\psi_i \approx \psi_i$ and $\zeta_j' = \tan\zeta_i \approx \zeta_i$ for large values of ω_β. Furthermore, under those circumstances, $\rho(\psi_j, \beta_i) \approx |\psi_j - \beta_i|$ and $\rho(\zeta_j, \beta_i) \approx |\zeta_j - \beta_i|$, i.e., the geodesic distance between the points in the manifold is very close to the Euclidean distance between their projections on the tangent space. Secondly, recall from Equation (2.4) that, as $\kappa_j \to \infty$, the link function G_{κ_j} we have chosen for the circular model will resemble the cumulative distribution of the normal distribution with variance $\frac{\pi^4}{2\kappa_j+1}$. As a consequence of these two features, if we let both $\omega_\beta \to \infty$ and $\kappa_j \to \infty$ while keeping $\frac{\pi^4\omega_\beta}{2\kappa_j+1} = 1$, the likelihood function for the spherical model will converge to the likelihood of a one-dimensional Euclidean model with a probit link constructed on the tangent space at 0. Furthermore, under these circumstances, the von Mises prior on the spherical coordinates maps onto the widely used Gaussian prior on the tangent space (recall Equation (2.6)).

The previous discussion suggests that we can use the variance of the ideal points to measure the level of circularity in the policy space of a given dataset. In particular small values for this variance indicate that the policy space is approximately Euclidean, and vice versa. We will make use of this observation in Section 2.4.3.

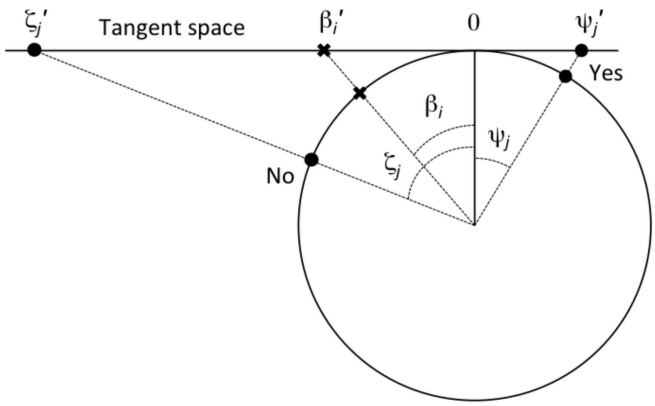


Figure 2.2: The circular manifold and its projection on the tangent space at the origin (located in our graphs at the upper pole). The values β_i, ψ_j and ζ_j correspond to the coordinates in the circular policy space (measured as angles with respect to vertical axis), while the values of β_i', ψ_j' and ζ_j' are their projections on the tangent space.

Another useful interpretation of our model that arises from this connection is as an interpolator between the one-dimensional and the two-dimensional Euclidean models. Indeed, in addition to the natural geometric argument that arises from embedding the circle into a two-dimensional Euclidean space, we note that the likelihood associated with the 1D Euclidean model involves $\mathcal{O}(I + 2J)$ parameters, the one for the circular model involves $\mathcal{O}(I + 3J)$ parameters, and the one for the two-dimensional Euclidean involves $\mathcal{O}(2I + 4J)$. This means that the circular model provides slightly more degrees of freedom to fit the data than a one-dimensional Euclidean model, but less than a two-dimensional Euclidean.

2.2.5 Ranking legislators under the circular model

The unit circle is not endowed with a total order, which represents a challenge if our goal is to rank legislators using the latent scale. We get around this issue by unwinding the circle into a traditional linear scale in $(-\pi, \pi]$. Breaking the circle at the bottom pole is natural if we consider the fact that the prior on the ideal points is centered at 0 (which corresponds to the middle of this interval), as well as the behavior of the prior when $\omega_\beta \to \infty$.

Unwinding the circle might seem somewhat ad-hoc after our heavy emphasis on the circular nature of the policy space. A formal justification is as follows: if the ideal points $\beta_1, \ldots, \beta_I$ all lie on the interval $(-\pi/2, \pi/2)$ (i.e., the upper semi-circle in Figure 2.1) the ranking of the projection of the ideal points on the tangent space at 0 (given by $\beta_i' = \tan \beta_i$, recall Figure 2.2) is identical to the ranking generated by unwinding the circle (since the tangent is a monotonic function when restricted to this domain). This scenario (all β_is in the $(-\pi/2, \pi/2)$ interval) is not a contrived one: it is an assumption that underlies the construction of our model, one that seems to be supported in most of the examples discussed in Section 2.4, including some of those in which there is evidence that the circular model dominates the Euclidean one.

2.3 Computation

The posterior distribution for the model, which takes the form

$$p(\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\kappa}, \omega_\beta, \lambda \mid \boldsymbol{Y}) \propto \left[\prod_{i=1}^{I}\prod_{j=1}^{J}\left\{G_{\kappa_j}\left(e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right)\right)\right\}^{y_{i,j}}\left\{1 - G_{\kappa_j}\left(e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right)\right)\right\}^{1-y_{i,j}}\right] \left[\prod_{i=1}^{I}\frac{\exp\left\{\omega_\beta \beta_i\right\}}{2\pi I_0(\omega_\beta)}\right]\left[\frac{1}{\theta}\exp\left\{-\frac{\omega_\beta}{\theta}\right\}\right]\left[\prod_{j=1}^{J}\frac{1}{\lambda}\exp\left\{-\frac{\kappa_j}{\lambda}\right\}\right]\left[\frac{\xi}{\lambda^2}\exp\left\{-\frac{\xi}{\lambda}\right\}\right], \quad (2.7)$$

is analytically intractable. Hence, inference for the model parameters is carried out using Markov chain Monte Carlo (MCMC) algorithms.

The algorithm we propose is a hybrid that combines Gibbs sampling, random walk Metropolis-Hastings and Hamiltonian Monte Carlo (HMC) steps to sample from the conditional distributions of each parameter. The simplest steps correspond to sampling the parameters λ, ω_β and $\kappa_1, \ldots, \kappa_J$. More specifically, we sample λ from its inverse-Gamma full conditional posterior distribution, and sample ω_β as well as each of the κ_js using random walk Metropolis Hastings with log-Gaussian proposals. The variance of these proposals are tuned so that the acceptance rate is roughly 40%. On the other hand, for sampling the latent positions we employ the Geodesic Hamiltonian Monte Carlo (GHMC) algorithm described in Section 1.3.3.

As an example, consider the step associated with updating each of the β_is. From Equation (2.7), the density (with respect to the Lebesgue measure on $[-\pi, \pi]$) of its full conditional distribution takes the form

$$p(\beta_i \mid \cdots) \propto \exp\{\omega_\beta \beta_i\}$$
$$\prod_{j=1}^{J} \left\{ G_{\kappa_j}\left(\{\arccos(\cos(\zeta_j - \beta_i))\}^2 - \{\arccos(\cos(\psi_j - \beta_i))\}^2 \right) \right\}^{y_{i,j}}$$
$$\left\{ 1 - G_{\kappa_j}\left(\{\arccos(\cos(\zeta_j - \beta_i))\}^2 - \{\arccos(\cos(\psi_j - \beta_i))\}^2 \right) \right\}^{1-y_{i,j}},$$

while the density of the associated Hausdorff measure in $\mathbb{R}^2$ is given by

$$p(\boldsymbol{x}_{\beta_i} \mid \cdots) \propto \exp\left\{ \boldsymbol{\eta}_\beta^T \boldsymbol{x}_{\beta_i} \right\}$$
$$\prod_{j=1}^{J} \left\{ G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right\}^{y_{i,j}}$$
$$\left\{ 1 - G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right\}^{1-y_{i,j}}, \qquad \boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1,$$

where $\boldsymbol{\eta}_\beta^T = (\omega_\beta, 0)$, $\boldsymbol{z}_{\psi_j}^T = (\cos\psi_j, \sin\psi_j)$, $\boldsymbol{z}_{\zeta_j}^T = (\cos\zeta_j, \sin\zeta_j)$, and the mapping from β_i to $\boldsymbol{x}_{\beta_i}$ is $\boldsymbol{x}_{\beta_i}^T = (\cos\beta_i, \sin\beta_i)^T$. Given tuning parameters ϵ (the step size) and L (the number of steps), the GHMC sampler takes the form:

1. Map the current value of the chain, $\beta_i^{(c)}$ onto the embedding space $\mathbb{R}^2$ by setting $\boldsymbol{x}_{\beta_i}^{(c)} = \left(\cos\beta_i^{(c)}, \sin\beta_i^{(c)}\right)^T$ and initialize $\boldsymbol{x}_{\beta_i} = \boldsymbol{x}_{\beta_i}^{(c)}$.

2. Initialize the auxiliary momentum variable $\boldsymbol{\phi}$ by sampling $\boldsymbol{\phi} \sim N(0, \boldsymbol{I}_2)$.

3. Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\phi} \leftarrow \left(\boldsymbol{I}_2 - \boldsymbol{x}_{\beta_i}\boldsymbol{x}_{\beta_i}^T\right)\boldsymbol{\phi}$, and then set $\boldsymbol{\phi}^{(c)} = \boldsymbol{\phi}$.

4. For each of the L leap steps:

   (a) Update the momentum by setting $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2}\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots)$.

   (b) Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\phi} \leftarrow \left(\boldsymbol{I}_2 - \boldsymbol{x}_{\beta_i}\boldsymbol{x}_{\beta_i}^T\right)\boldsymbol{\phi}$, and then set the angular velocity of the geodesic flow

$\nu = \|\phi\|$.

(c) Update $\boldsymbol{x}_{\beta_i}$ and $\boldsymbol{\phi}$ jointly according to the geodesic flow with step size of ϵ,

$$\boldsymbol{x}_{\beta_i} \leftarrow \boldsymbol{x}_{\beta_i} \cos(\nu\epsilon) + \frac{\phi}{\nu} \sin(\nu\epsilon), \qquad \phi \leftarrow \phi \cos(\nu\epsilon) - \nu\, \boldsymbol{x}_{\beta_i} \sin(\nu\epsilon).$$

(d) Update $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{\epsilon}{2} \nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots)$.

(e) Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\phi} \leftarrow \left(\boldsymbol{I}_2 - \boldsymbol{x}_{\beta_i} \boldsymbol{x}_{\beta_i}^T\right) \boldsymbol{\phi}$.

5. Set the proposed values as $\boldsymbol{x}_{\beta_i}^{(p)} = \boldsymbol{x}_{\beta_i}$, $\boldsymbol{\phi}^{(p)} = \boldsymbol{\phi}$, and $\beta^{(p)} = \arctan2\left(x_{\beta_i,2}, x_{\beta_i,1}\right)$. The proposed value $\beta^{(p)}$ is accepted with probability

$$\min\left\{1, \frac{p_{\mathcal{H}}\left(\boldsymbol{x}_{\beta_i}^{(p)} \mid \cdots\right) \exp\left\{-\frac{1}{2}\left[\boldsymbol{\phi}^{(p)}\right]^T \boldsymbol{\phi}^{(p)}\right\}}{p_{\mathcal{H}}\left(\boldsymbol{x}_{\beta_i}^{(c)} \mid \cdots\right) \exp\left\{-\frac{1}{2}\left[\boldsymbol{\phi}^{(c)}\right]^T \boldsymbol{\phi}^{(c)}\right\}}\right\}$$

Detailed expressions for the Hausdorff measures associated with the full conditional distributions of the β_is, ψ_js and ζ_js, as well as their gradients, can be seen in Appendix A.1. In our implementation of the algorithm, we periodically vary the value of the tuning parameters ϵ and L by randomly sampling them from predetermined distributions. These changes are done independently of the current value of the parameter, thereby preserving the Markovian structure of the algorithm. This approach, sometimes called "jittering" in the literature (e.g., see 77, pg. 306), greatly improves the mixing of the algorithm in our experiments. The specific range in which ϵ and L move for each parameter and dataset is selected to target an average acceptance probability between 60% and 90% [65, 67].

2.4 Circular voting in the modern U.S. Congress

In this Section we analyze roll call voting data from the modern U.S. House of Representatives. We first present legislator-level results for two specific Houses (the 116^{th} and the 112^{th}), and then show a longitudinal analysis of chamber-level summaries covering the 100^{th} to the 116^{th} Houses. In all these analyses, the number of leaps used in the HMC steps is randomly selected from a discrete uniform distribution between 1 and 10 every 50 samples. Similarly, the leap sizes are drawn from uniform distribution on $(0.01, 0.04)$ or $(0.005, 0.04)$ for each β_i, and from a uniform distribution on $(0.01, 0.105)$ for each ζ_j and ψ_j. All inference presented in this Section are based on 20,000 samples obtained after convergence of the Markov chain Monte Carlo algorithm. The length of the burn in period varied between 20,000 and 80,000 iterations depending on the dataset, with a median around 30,000. Convergence was checked by monitoring the value of the log-likelihood function, both through visual inspection of the trace plot, and by comparing multiple chains using the procedure in [78]. Details on the convergence and mixing analysis for the 116^{th} House and the 112^{th} House can be seen in its corresponding section. For each House, we excluded legislators who are absent for more than 40% of the votes.

2.4.1 The Squad, revisited

First, we revisit the voting record of the first session of the 116^{th} Congress discussed in Section 2.1. Figure 2.3 presents trace plots for the log-likelihood of the model for two runs (panel (a) corresponds to the run on which the results in this section are based, while panel (b) contains a second run obtained from a different,

randomly selected, starting point). Both runs include a total of 100,000 samples, and the vertical line indicates the end of the burn-in period. Convergence of the two chains to a common mode was checked using the Gelman and Rubin test [78] (observed potential scale reduction factor is 1.012, with an upper limit of 1.058 for the associated 95% confidence interval). In this dataset, we can see that the algorithm quickly moves close to a high probability area, but once there, takes a while to find the mode. To further emphasize that the two runs identify the same mode, we present in Figure 2.4 a comparison of the rank order of legislators generated from both runs. We can see that the two sets of results are nearly identical, with the very small differences that can be observed likely being the result of Monte Carlo error.

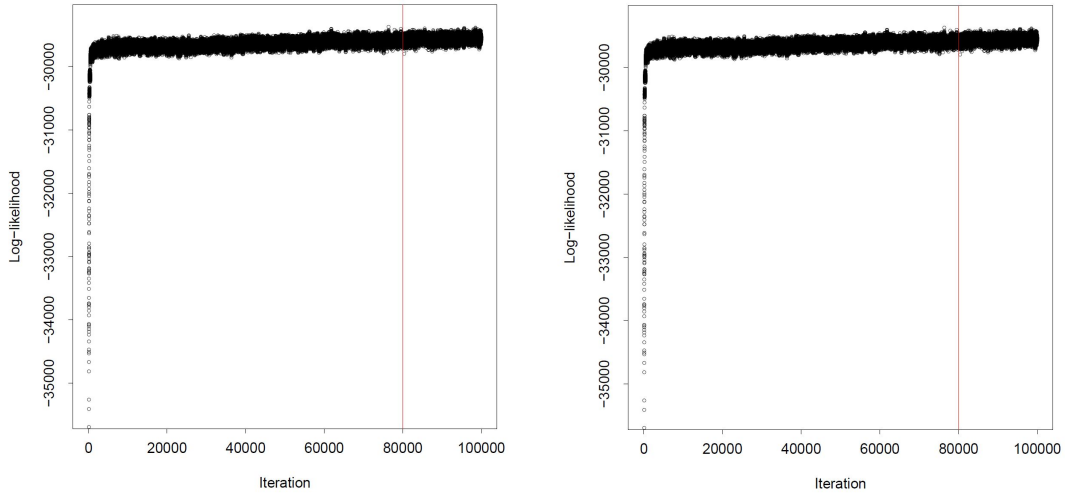


Figure 2.3: Trace plots for the log-likelihood associated with two runs of the MCMC algorithm for the roll call data from the first session of the 116$^{\text{th}}$ U.S. House of Representatives.

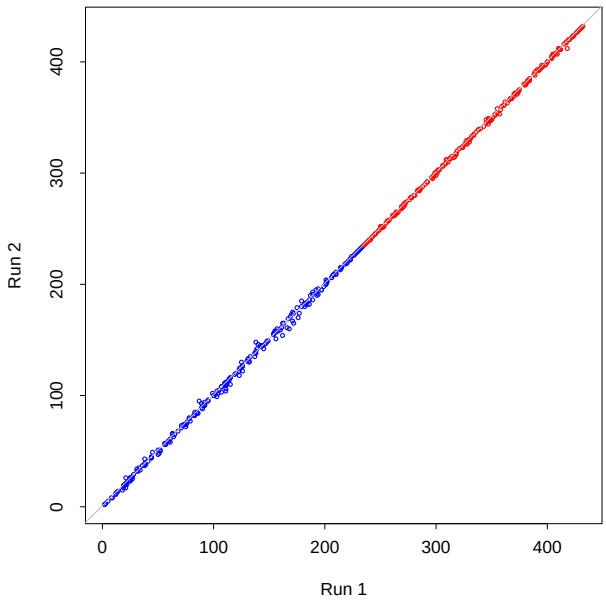


Figure 2.4: Comparison of the posterior median ranks of the legislators obtained from each of the two runs for the 116$^{\text{th}}$ U.S. House of Representatives.

Figure 2.5 shows the effective sample sizes associated with the rank orders generated by each of the two runs. We focus on the rank orders rather than the β_is because (1) these are the key quantities of interest in our analysis, and (2) the ranks are identifiable from the data. Furthermore, because of the large number of legislators involved (432), the results are presented in the form of boxplots. For most legislators, the effective sample size is quite reasonable (75% of the legislator's ranks have an effective sample size above 2,000), particularly when considering the complexity of the model. However, we do see that there is a small number of legislators for which the effective sample size is very low. This behavior, which might seem alarming at first sight, reflects a limitation of the effective sample size as a measure of mixing, and not an issue with our algorithm and results. Indeed, note that the ranks are are discrete parameters, and that they are not independent across legislators. This means that, even if the algorithm

is properly exploring the parameter space, the implied ranks might remain the same over multiple iterations. We found this situation arising for a couple of very extreme legislators. The prototypical example is Justin Amash, who is estimated by our model to be the most extreme Republican in the 116$^{\text{th}}$, with no uncertainty associated with such a rank (see Table 2.3). For Amash, the posterior samples of the rank are highly autocorrelated, leading to a low effective sample size even if the algorithm is mixing properly.

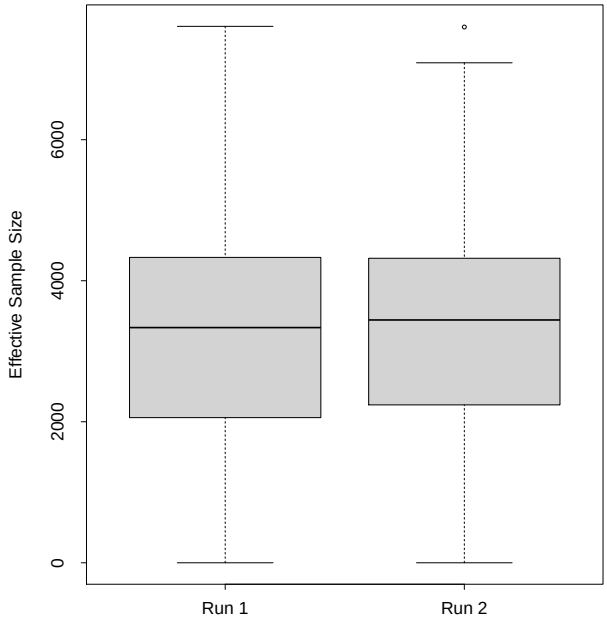


Figure 2.5: Effective Sample Size of the rank order of legislators generated by the two runs for the 116$^{\text{th}}$ U.S. House of Representatives.

Table 2.2 reports the posterior median rank order and associated 95% credible intervals for the members of the Squad according to our circular model. The difference between these and those we reported in Table 2.1 is dramatic, with the circular model clearly placing Presley, Omar, Tlaib and Ocasio-Cortez among the most liberal members of the Democratic party.

Table 2.2: Median rank of the members of the "Squad" during the first session of the 116th U.S. House of Representatives according to our circular model. Lower numbers for the ranks correspond to more liberal legislators. Numbers in parenthesis correspond to 95% credible intervals.

	Rank Order (Circular)
Pressley (D MA-7)	5 (1,21)
Omar (D MN-5)	2 (1,8)
Tlaib (D MI-13)	2 (1,9)
Ocasio-Cortez (D NY-14)	3 (1,11)

Table 2.3: Median rank of three selected Republican legislators during the first session of the 116th U.S. House of Representatives according to two models: a one-dimensional Euclidean voting model, and our circular model. Higher numbers for the ranks correspond to more conservative legislators. Numbers in parenthesis correspond to 95% credible intervals.

	Rank Order	
	Euclidean (1D)	Circular
Amash (R MI-3)	249 (244,255)	432 (432,432)
Massie (R, KY-4)	356 (334,375)	431 (430,431)
Gaetz (R FL-1)	314 (297,332)	418 (412,423)

More generally, Figure 2.6 compares the rank order of legislators estimated using the one-dimensional Euclidean model to the rank order estimated by the circular model. On the Democratic side, we can see some substantial differences in the ranks estimated by the Euclidean model versus those estimated by the circular model. However, it is clear that the largest differences correspond to the four members of the Squad. In contrast, on the Republican side, the ranks estimated by both models are generally in close agreement. The three main exceptions are

representatives Justin Amash (MI-3), Thomas Massie (KY-4) and Matt Gaetz (FL-1), who are estimated to be more extreme by the circular model (see also Table 2.3). An inspection of their record suggests that the ranking generated by the circular model is quite sensible. In particular, consider Justin Amash and Thomas Massie. Justin Amash is a libertarian-leaning conservative first elected in 2010 as a Republican. He has received high scores from from right-leaning interest groups such as the Club for Growth, Heritage Action for America, and Americans for Prosperity, and praise from conservative think tanks and nonprofit organizations. He was also a founding member of the House Freedom Caucus, a group of hard-line conservative Republicans in the House of Representatives. However, he is also widely known for his contrarian views and for voting with Democrats in certain issues. For example, he was the only Republican to vote against the "In God We Trust" House Resolution passed in November 2011 and the House Resolution supporting the officers and personnel of Immigration and Customs Enforcement (ICE) in July 2018. Furthermore, he co-sponsored a bill by Democrat Ayanna Pressley (one of the members of the Squad) that would abolish the death penalty at the federal level. In fact, Amash left the Republican party in July 2019 to become an independent, and became the only non-Democrat in the House to vote in favor of an impeachment inquiry into the activities of President Trump and of either of the two articles of impeachment. Thomas Massie is another libertarian leaning Republican representative often associated with the House Liberty Caucus of Tea Party Republicans. However, he is also know for often being the only member of the House to vote against a number of resolutions. For example, on March 27, 2020, Massie forced the return to Washington of members of the House (who were sheltering in place in the midst of the Covid-19 crisis) by withholding unanimous consent on the passage of the The Coronavirus

Aid, Relief, and Economic Security Act (CARES) Act.

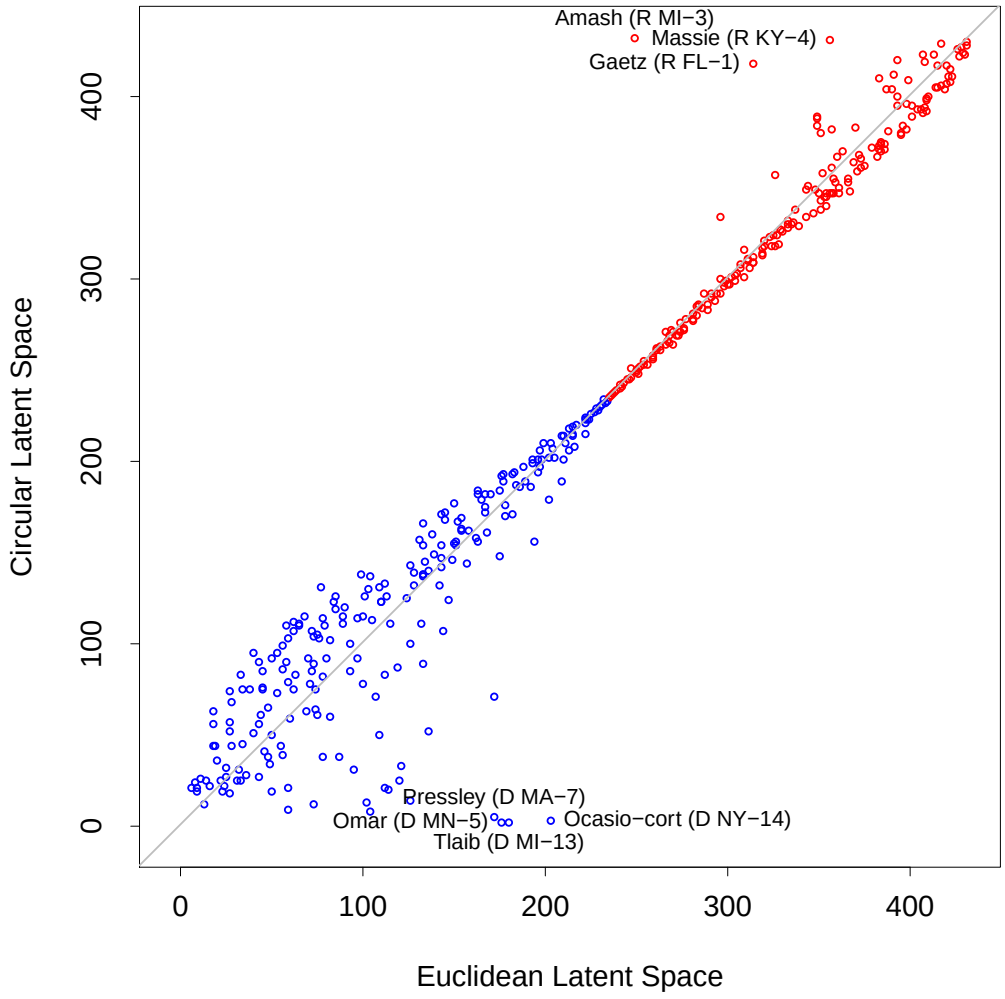


Figure 2.6: Posterior median of the rank-order in the first session of the 116$^{\text{th}}$ U.S. House of Representatives under the one-dimensional Euclidean (horizontal axis) and the circular (vertical axis) models.

To complete this illustration, we provide specific information about various bills in which both circular and Euclidean voting patterns are present. First, Figure 2.7 provides two examples of circular voting in the 116$^{\text{th}}$ House of Representatives. The first one, HRES246, opposed the global boycott, divestment, and sanctions

movement, as well as other efforts targeting Israel. This resolution (1) urged both sides in the Israel-Palestinian conflict to return to direct negotiations, (2) expressed support for a solution resulting in the state of Israel existing alongside a democratic Palestinian state, and (3) reaffirmed the right of U.S. citizens to free speech, including the right to protest or criticize U.S. or foreign government policies. HRES246 was opposed by a group of 16 Democrats (including three members of the Squad), as well as by Representative Massie. Massie opposed the measure because it "calls for Israel to implement a so-called two-state solution. Rather than dictate to Israel what the U.S. believes is best for Israel, Congress should instead refrain from interfering with Israel's own decisions regarding its foreign and domestic policy." He also stated that he "do[es] not support federal efforts to condemn any type of private boycott, regardless of whether or not a boycott is based upon bad motives. These are matters that Congress should properly leave to the States and to the people to decide." Both of these are traditional "libertarian" arguments. On the other hand, the main driver for Democrats voting against this resolution was support for Palestine. For example, in her floor speech, representative Tlaib invoked her Palestinian grandmother in opposing the resolution, which she said "attempts to delegitimize a certain people's political speech and send a message that our government can and will take action against speech it doesn't like." While there seems to be some ideological common ground between both positions (in particular, a shared desire to limit government impingement on free speech), it is clear that the underlying ideology is completely different, making this an instantiation of the Horseshoe Theory in the context of the U.S. House of Representatives.

The second example in this category is S1790, the National Defense Authorization Act for Fiscal Year 2020. S1790 authorized FY2020 appropriations and set

forth policies for Department of Defense (DOD) programs and activities, including military personnel strengths. S1790 was opposed by two Democrats (Tlaib, who is one of the members of the squad, and Blumenauer), as well as by two Republicans (Massie and Amash, whom we have already discussed). Note that, for both HRES246 and S1790, the "Nay" position is roughly located opposite to the (circular) average of all ideal points, while the "Yea" position is located close to the (circular) average of the ideal points of the legislators that voted in favor of the measure. Furthermore, the posterior distribution of κ_j indicates moderate to low concentration values for the link function for this kind of votes.

On the other hand, Figure 2.8 shows two examples of Euclidean voting, HRES5 and HR135. HRES5, which sets forth the rule for consideration of HRES6 (adopting the Rules of the House of Representatives for the 116th Congress), HR21 (Consolidated Appropriations Act, 2019), and HJRES1 (FY2019 Department of Homeland Security appropriations), was voted strictly along party party lines. We see in this case that the "Yea" and "Nay" positions are located at either side of the parties, and that the posterior distribution ofκ_j favors relatively large concentration values. This configuration is very similar to the one that is obtained by fitting a Euclidean model to the data. On the other hand, HR135, the Elijah E. Cummings Federal Employee Antidiscrimination Act of 2019, requires each federal agency to establish a model Equal Employment Opportunity Program that is independent of the agency's Human Capital or General Counsel office, and it establishes requirements related to complaints of discrimination and retaliation in the workplace. HR135 was voted unanimously (except for 8 abstentions). Note that the "Yea" and "Nay" positions in this case are similar to those estimated for the circular votes, but the corresponding value of κ_j is much higher for the fully unanimous vote.

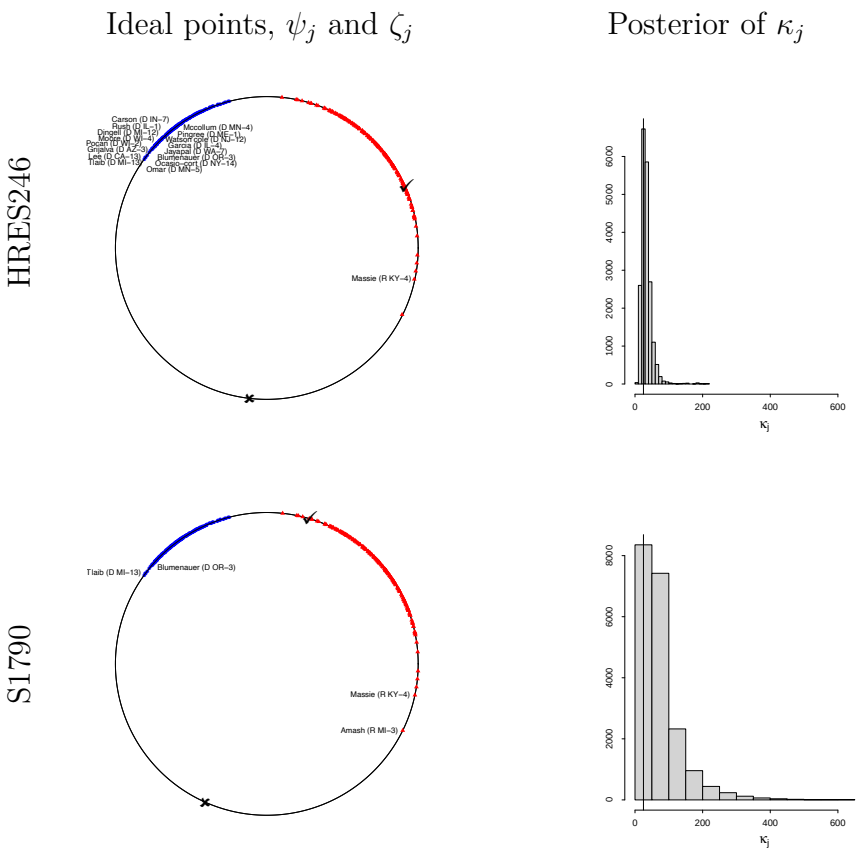


Figure 2.7: Two examples of circular voting patterns during the 116$^{\text{th}}$ House of Representatives. Graphs on the left column depict the posterior mean ideal point for the legislators (which are the same on both plots), along with the "Yea" and "Nay" positions (represented through a check mark and a cross, respectively). The names in the graphs correspond to the legislators that voted against the measure. The right columns presents a histogram of samples of the posterior distribution of the corresponding κ_j. The vertical line corresponds to $\mathrm{E}(1/\lambda \mid \text{data}) = \mathrm{E}(\kappa_j \mid \text{data})$.

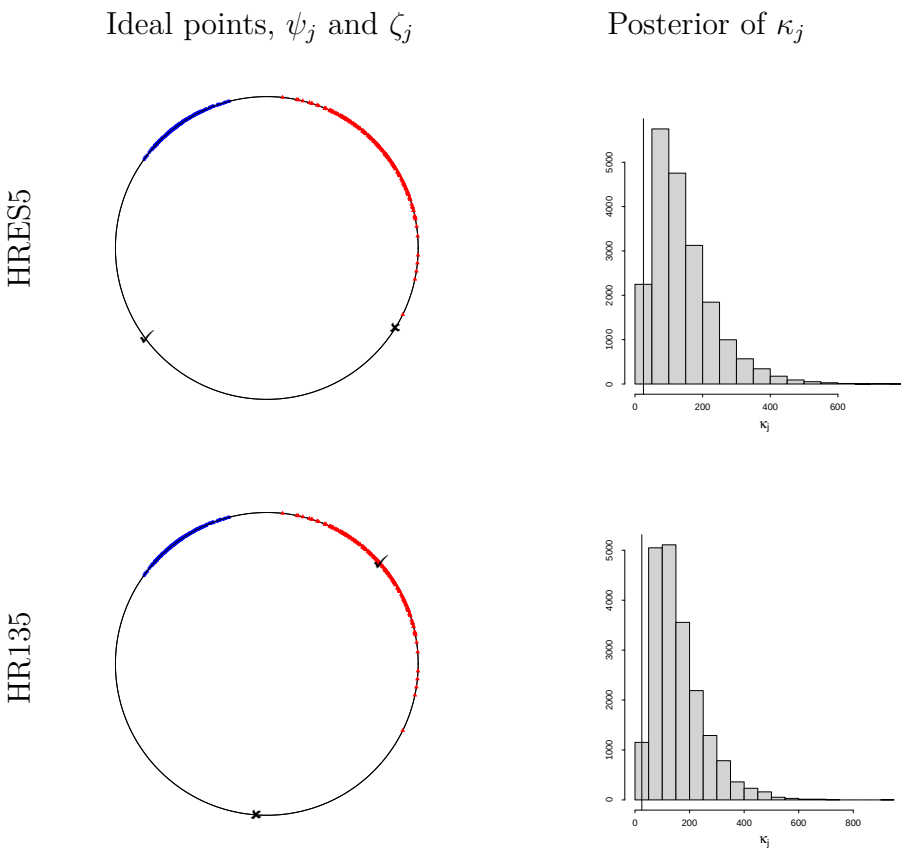


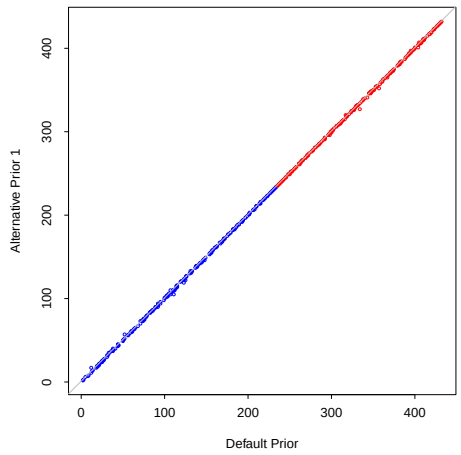
Figure 2.8: Two examples of Euclidean voting patterns during the 116$^{\text{th}}$ House of Representatives. Graphs on the left column depict the posterior mean ideal point for the legislators (which are the same on both plots), along with the "Yea" and "Nay" positions (represented through a check mark and a cross, respectively). The right columns presents a histogram of samples of the posterior distribution of the corresponding κ_j. The vertical line corresponds to $\mathrm{E}(1/\lambda \mid \text{data}) = \mathrm{E}(\kappa_j \mid \text{data})$.

Sensitivity analysis

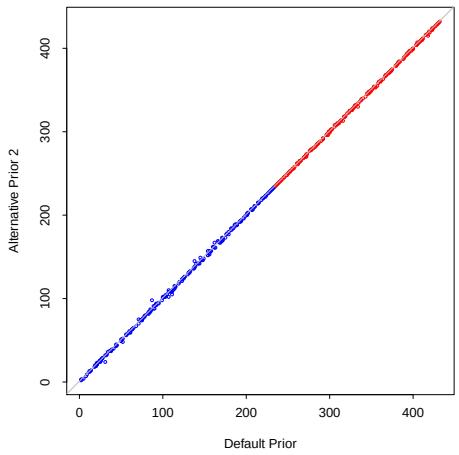
In order to understand the effect of the priors on our inferences, we conducted a sensitivity analysis by refitting the model under two alternative priors for the 116$^{\text{th}}$ Houses. First, we consider a Gamma prior with shape parameter $a = 7$ and rate parameter $b = 1$ for the β_is (so that $\Pr(-\pi/2 < \beta_i < \pi/2) \approx 0.995$ a priori)

and set $\xi = 100$. Relatively speaking, this prior favors configurations that are closer to one-dimensional Euclidean model. Secondly, we consider an exponential prior with rate parameter $\theta = 2$ for the β_is (so that $\Pr(-\pi/2 < \beta_i < \pi/2) \approx 0.80$ a priori) and set $\xi = 25$. Compared to the first hyperprior, this second one favors circular configurations.

Figures 2.9 compare the rank order of the legislators generated under our default prior with each of our two alternative priors for the 116$^{\text{th}}$ Houses. In both cases, we can see that the results are nearly identical in all cases, with the very small differences observed in the graphs being likely driven by Monte Carlo noise.



(a) Alternative prior 1 (favors 1D Euclidean model)



(b) Alternative prior 2 (favors circularity)

Figure 2.9: Posterior median rank comparison between default prior and both alternative priors for the 116$^{\text{th}}$ House.

2.4.2 The Conservative Revolt of 2010

The election in November 2008 of Barack Obama as president of the United States generated a strong conservative backlash that has had a profound impact on U.S.

politics in general, and on the Republican party in particular [38]. This backlash influenced the results of the 2010 midterm election [36]. The 112$^{\text{th}}$ Congress had a large Republican majority (in fact, had the largest Republican majority since the 80$^{\text{th}}$ Congress in the late 1940s). It was also the first Congress in over 150 years in which the Republican party held the House but not the Senate, and the first Congress to begin with the House and the Senate controlled by different parties since the 99$^{\text{th}}$ Congress (1985-1987).

Among the 242 Republican legislators elected to the 112$^{\text{th}}$ House of Representatives was a large group of insurgent candidates, many of them backed by a loose grassroots coalition ostensibly built on the principles of fiscal responsibility, adherence to the Constitution, and limited government, that has become known as the Tea Party movement [37]. Many of these insurgent legislators went on to form congressional member organizations such as the Tea Party caucus and the House Liberty caucus (both founded during the 112$^{\text{th}}$ Congress, the first in July 2010 and the second in March 2011), as well as the House Freedom caucus (founded in 2015 during the 114$^{\text{th}}$ Congress). These three caucuses are all considered to represent the most extreme wing of the Republican party, and some recent evidence suggests that their members vote like a significantly farther-right third party in Congress (e.g., see 79). However, as we will see shortly, many of their members are ranked by traditional spatial voting models as mainstream, or even centrist Republicans.

Similarly to the previous section, we present in Figure 2.10 trace plots for the log-likelihood of the model for two runs (panel (a) corresponds to the run on which the results in this section are based, while panel (b) contains a second run obtained from a different, randomly selected, starting point). Both runs include a total of 100,000 samples, and the vertical line indicates the end of the burn-in period. As

before, convergence of the two chains to a common mode was checked using the Gelman and Rubin test (in this case, the observed potential scale reduction factor is 1.006, with an upper limit of 1.029 for the associated 95% confidence interval). We again observe that the algorithm quickly moves close to a high probability area, but once there, takes a while to find the mode. Finally, Figure 2.11 presents a comparison of the rank order of legislators generated from both runs. As before, the two sets of results are near identical, with the very small differences that can be observed likely being the result of Monte Carlo error.

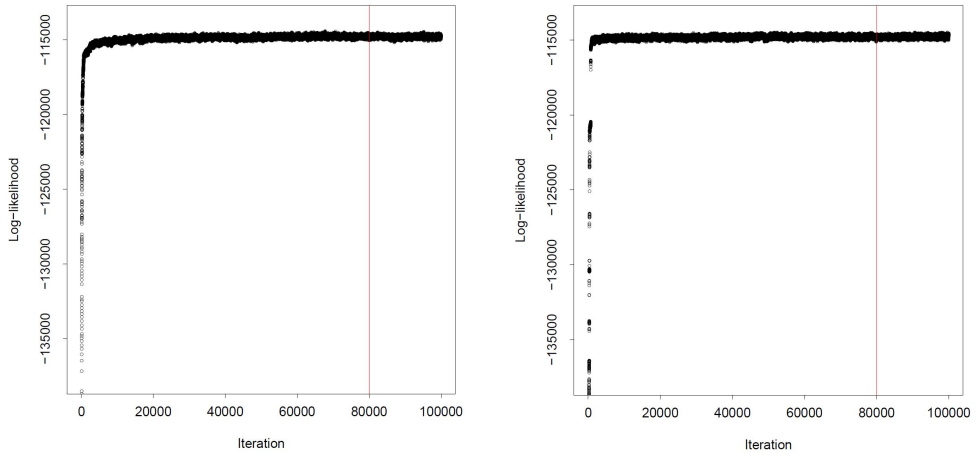


Figure 2.10: Trace plots for the log-likelihood associated with two runs of the MCMC algorithm for the roll call data from the 112$^{\text{th}}$ U.S. House of Representatives.

As in Figure 2.5, Figure 2.12 shows the effective sample sizes associated with the rank orders generated by each of the two runs. Overall, the effective sample sizes for the data from the 112$^{\text{th}}$ House tend to be somewhat lower than in the case of the of the 116$^{\text{th}}$. However, the values still seem reasonable (75% of the legislator's ranks have effective sample sizes above 1,200). We again see a few ranks with quite low effective sample sizes, which we still attribute to the fact that ranks are discrete quantities that are not independent across legislators.

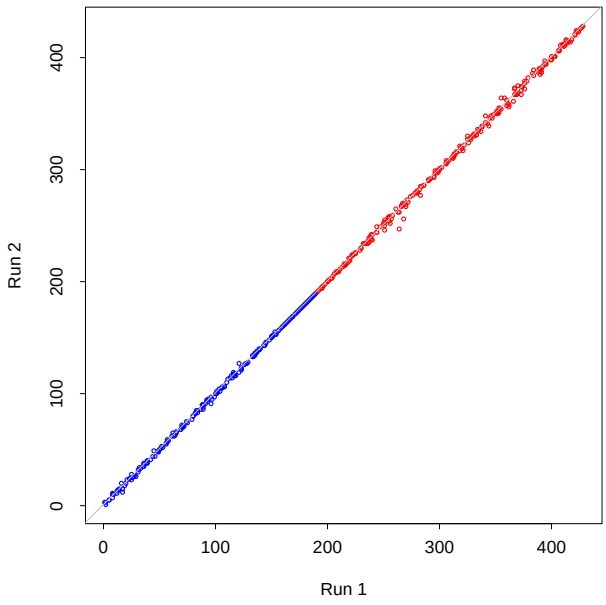


Figure 2.11: Comparison of the posterior median ranks of the legislators obtained from each of the two runs for the 112$^{\text{th}}$ U.S. House of Representatives.

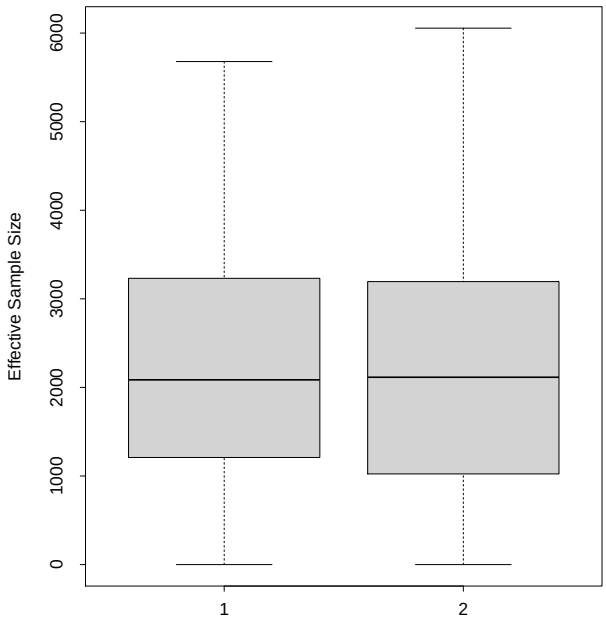


Figure 2.12: Effective Sample Size of the rank order of legislators generated by the two runs for the 112$^{\text{th}}$ U.S. House of Representatives.

Figure 2.13 compares the rank order of legislators estimated using a one dimensional model to the rank order estimated by our circular model. For the Democratic party, the ranking generated by both models are reasonably similar. The main outlier is Dennis Kucinich (OH-10), who is ranked as much more liberal by the circular model: his posterior median rank is 2 under the circular model, with a 95% credible interval of (1,6), but it is 76 under the Euclidean model, with a 95% credible interval (62,97). This more extreme ranking fits better with the widely-held perception that Kucinich was was one of the most liberal members of the United States House of Representatives during this period. In contrast, on the Republican side, we see some very large differences between the rankings generated by both models. In particular, we see a large group of legislators that are ranked as much more conservative by the circular model. Figure 2.15 provides additional details for the 15 Republican legislators for whom the difference in posterior median rankings between the one-dimensional Euclidean and the circular models is largest. It is interesting that, in all cases, the ranks assigned by the circular model are more extreme, and that 12 out of the 15 legislators in this list either were members of the Tea Party or the Liberty caucuses during this Congress, or later joined the Freedom caucus when it was formed. In particular, we must highlight that classifying Ron Paul, Justin Amash (who we already discussed in the previous section) or Jimmy Duncan as centrist (which is the implication from their rank under the Euclidean model) would be very hard to justify based on their stated political positions.

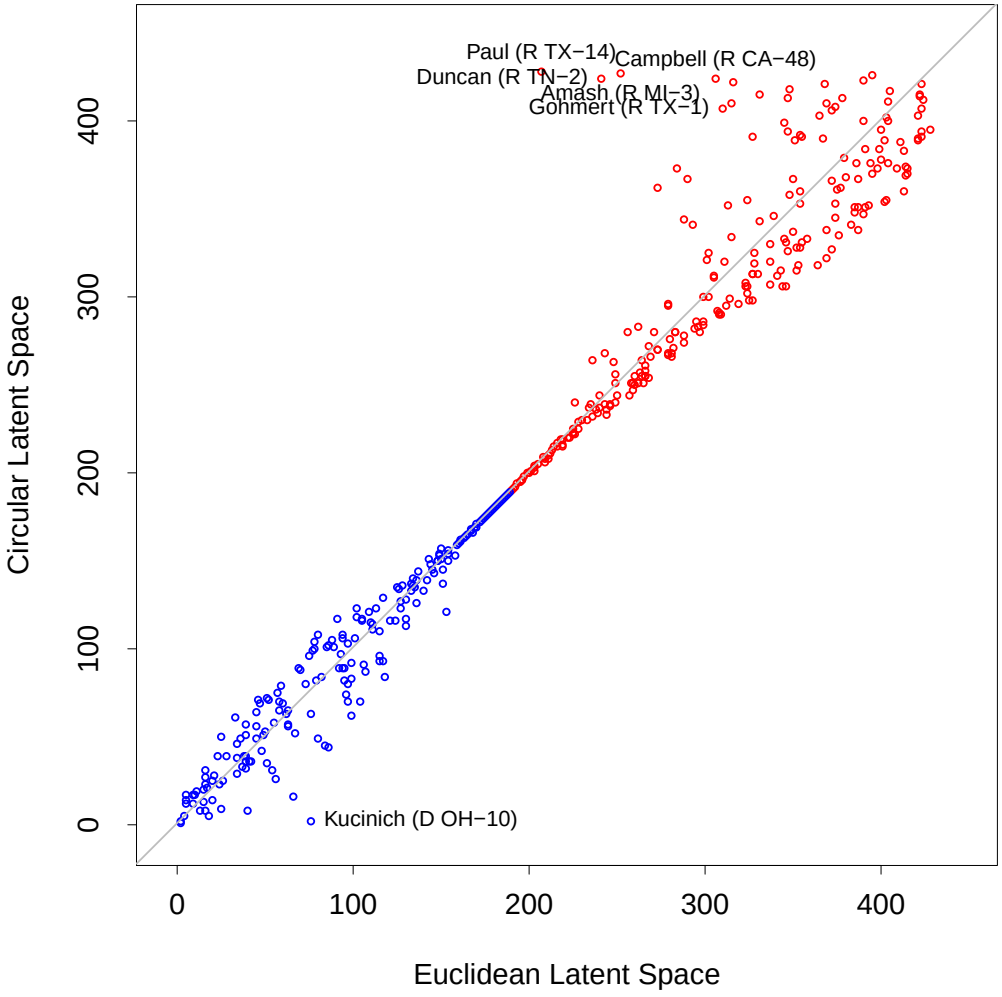
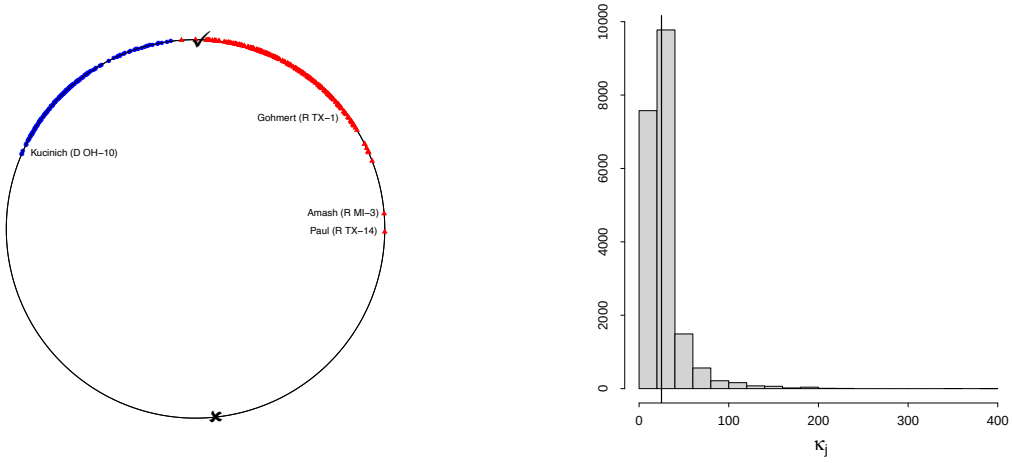


Figure 2.13: Posterior median of the rank-order in the 112$^{\text{th}}$ U.S. House of Representatives under the one-dimensional Euclidean (horizontal axis) and the circular (vertical axis) models.

Finally, Figure 2.14 presents one example of circular voting in the 112$^{\text{th}}$ House of Representatives. HR915 is the Jaime Zapata Border Enforcement Security Task Force Act, which amended the Homeland Security Act of 2002 to establish the Border Enforcement Security Task Force (BEST). It aimed at facilitating collaboration among federal, state, local, tribal, and foreign law enforcement agencies

to execute coordinated activities in furtherance of border security and homeland security, as well as to enhance information-sharing, including the dissemination of homeland security information among such agencies. Note that HR915 was opposed by representative Kucinich (according to our model, the most extreme Democrat in the House at the time) as well as by both Amash and Massie (in turn, the most extreme Republicans in the House according to our model) and by representative Louie Gohmert (whose ranking also significantly shifts under the circular model). As in Figure 2.7, the "Nay" position is roughly located opposite to the (circular) average of all ideal points, the "Yea" position is located close to the (circular) average of the ideal points of the legislators that voted in favor of the measure, and the posterior distribution of κ_j indicates moderate to low concentration values for the link function.



(a) Ideal points, ψ_j and ζ_j **(b)** Posterior of κ_j

Figure 2.14: Posterior median rank comparison between default prior and both alternative priors for the 112$^{\text{th}}$ House.

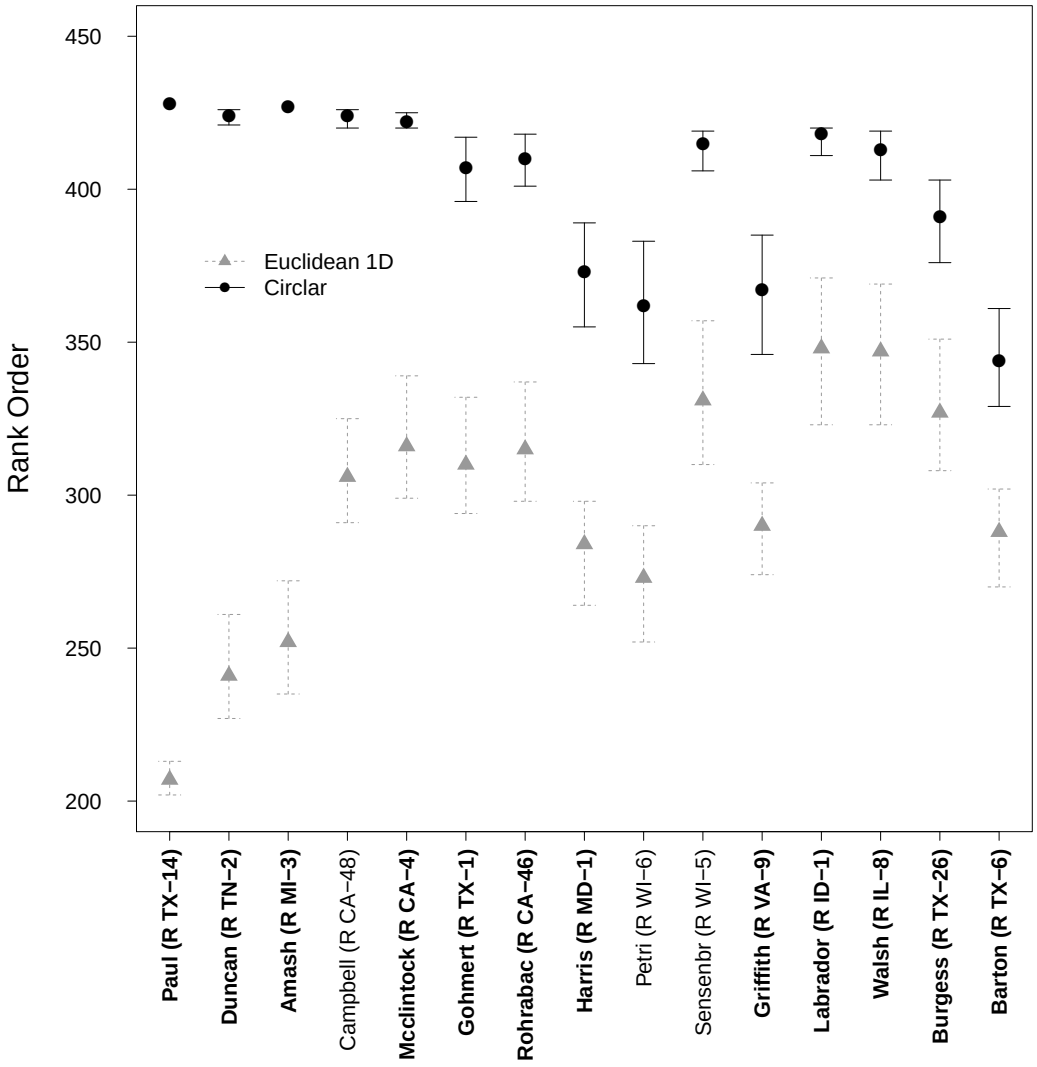
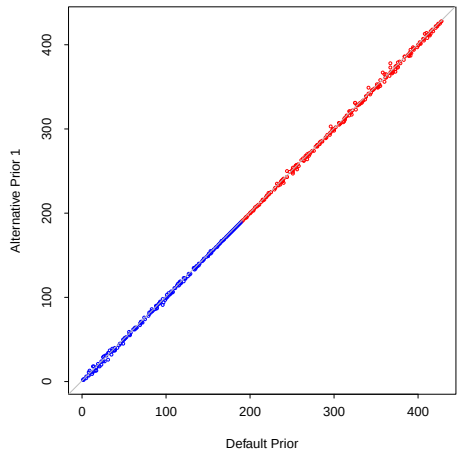


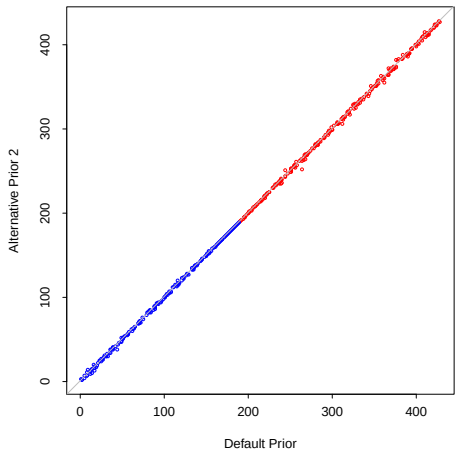
Figure 2.15: Posterior median ranks and associated 95% credible intervals for the fifteen Republican legislators in the 112$^{\text{th}}$ House for whom the difference in posterior median rankings between the one-dimensional Euclidean and the circular models is largest. Bolded names indicate that the legislator was a member of either the Liberty Caucus, the Freedom Caucus or the Tea Party Caucus at some of point of their career.

Sensitivity analysis

Similar to section 2.4.1, we conducted a sensitivity analysis by refitting the model under the same two alternative priors for the 112$^{\text{th}}$ Houses. Figures 2.16 compare the rank order of the legislators generated under our default prior with each of our two alternative priors for the 112$^{\text{th}}$ Houses. In both cases, we can again see that the results are nearly identical in all cases, with the very small differences observed in the graphs being likely driven by Monte Carlo noise.



(a) Alternative prior 1 (favors 1D Euclidean model)



(b) Alternative prior 2 (favors circularity)

Figure 2.16: Posterior median rank comparison between default prior and both alternative priors for the 112$^{\text{th}}$ House.

2.4.3 A longitudinal analysis of the contemporary U.S. House of Representatives

The previous two sections presented two very recent examples of circular voting behavior in the U.S. House of Representatives. We are interested now in understanding how pervasive this behavior has been in modern history. As we mention

in Section 2.2.4, the (circular) variance of the ideal points

$$\chi_0 = 1 - \left\{ \left(\frac{1}{I} \sum_{i=1}^{I} \cos \beta_i \right)^2 + \left(\frac{1}{I} \sum_{i=1}^{I} \sin \beta_i \right)^2 \right\}^{\frac{1}{2}}$$

provides a natural metric to measure the "circularity" of voting on a given Congress. It is important to note that this metric is useful in this context because, unlike the analogous metric for Euclidean models, it is comparable across Congresses (recall that the utility functions that underlie our formulation are not invariant to rescalings of the policy space).

Figure 2.17 shows the posterior mean and 95% credible intervals for χ_0 between the 100^{th} House (1987-1988) and (the first session of) the 116^{th} House (2019). We start the analysis from the mid 1980s because this period sits comfortably after the reforms of the mid-1970s (which included the introduction of electronic voting, leading to a dramatic increase in the number and nature of roll call votes recorded in the chamber). Note that χ_0 was relatively stable in the late 1980s and early 1990s, but then jumped when the control of the chamber switched from Democrats to Republican with the election of the 104^{th} House. It then remained more or less stable during the later half of the 1990s and the 2000s, to then fall during the 111^{th} House and then jump up again to historically high levels from the 112^{th} House on. While the overall increasing trend on χ_0 agrees well known patterns of increasing polarization in Congress, these results also suggest that such an increase in polarization has been accompanied by an increase in the frequency of "extremes voting together", a phenomenon that has not yet been fully documented or explored in the modern U.S. Congress .

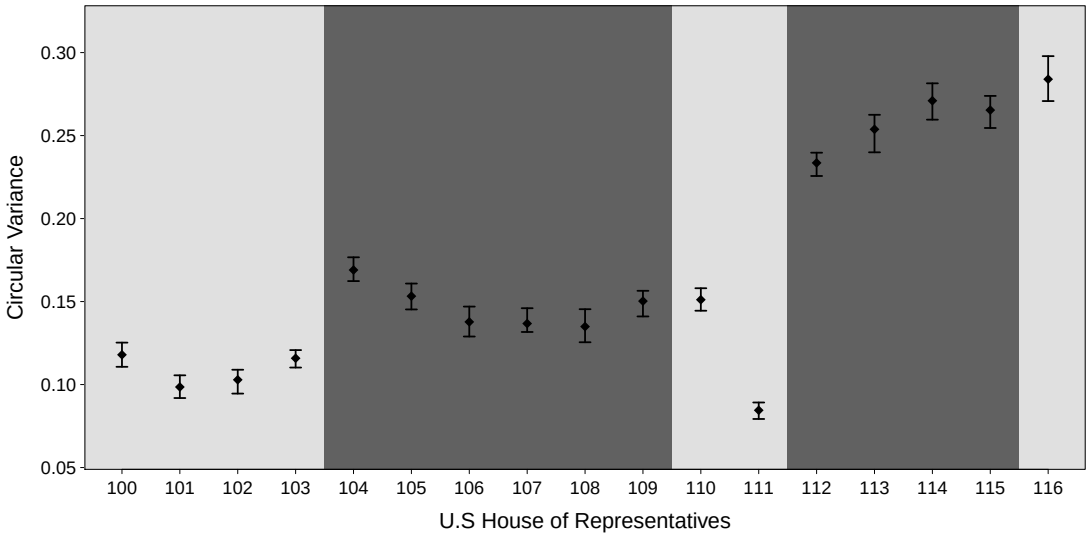


Figure 2.17: Circularity χ_0 for the 100^{th} to the 116^{th} U.S House of Representatives. Light gray background indicates a Democratic majority, while dark gray indicates a Republican majority. Results for the 116^{th} House include only the first session.

In order to better understand how circular voting has affected each party, we present in Figure 2.18 the circular variance associated with the ideal of points of both Democrat and Republican legislators, χ_D and χ_R, respectively. While some of the fluctuations in these metrics roughly match those we observed in Figure 2.17, it is clear that the overall trend has been for χ_D to decrease and for χ_R to increase over time. The divergence is particularly stark after the 112^{th}, with the Republican party showing an all-time high level of intraparty disagreement. Put another way, these results suggest that the Republican party has steadily become more fractious while the Democratic party has tended to unify, particularly over the last 10 years. This pattern is consistent with well-known political processes. On one hand, the steady decrease in χ_D between the 100^{th} and the 107^{th} Houses might be seen as the upshot of the the migration of the remaining former Southern Democrats to the Republican party. On the other hand, the large increase in χ_R

starting in the 112$^{\text{th}}$ can be understood as a consequence of the rise of the Tea Party movement (recall our discussion in Section 2.4.2).

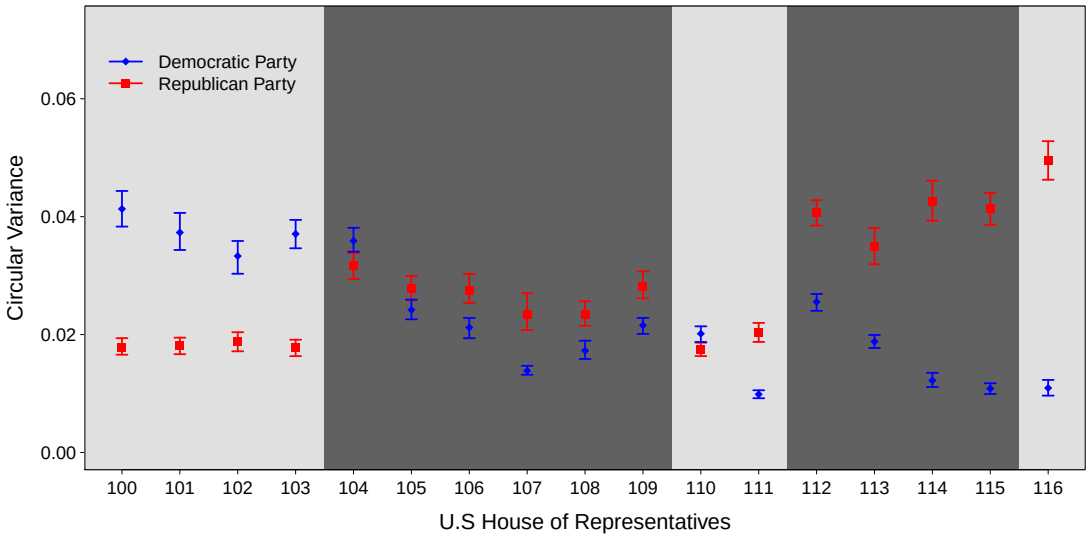


Figure 2.18: Within-party circular variances, χ_D and χ_R, for the 100$^{\text{th}}$ to the 116$^{\text{th}}$ U.S House of Representatives. Light gray background indicates a Democratic majority, while dark gray indicates a Republican majority. Results for the 116$^{\text{th}}$ House include only the first session.

To conclude this Section, we present in Table 2.4 the value of the Deviance Information Criteria (DIC, see [80, 81, 82]) associated with both the circular and one-dimensional Euclidean models fitted to the data from 100$^{\text{th}}$ to 116$^{\text{th}}$ U.S. House of Representatives. Similarly to the well-known Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC), the DIC balances goodness of fit against model complexity. However, unlike the AIC and the BIC, the DIC is well suited for hierarchical models where the number of effective parameters can be much smaller than the headline number. In our models, the DIC is defined as:

$$DIC = \ell\left(\mathrm{E}\{\boldsymbol{\Theta} \mid \text{data}\}\right) - 2\mathrm{Var}\left(\ell\left(\boldsymbol{\Theta}\right) \mid \text{data}\right),$$

where $\mathbf{\Theta} = [\theta_{i,j}]$ is the matrix of probabilities given by $\theta_{i,j} = \left\{G_{\kappa_j}\left(e_{i,j}\left(\psi_j, \zeta_j, \beta_i\right)\right)\right\}$ for the spherical model and $\theta_{i,j} = \Phi\left(\mu_j + \alpha_j\beta_i\right)$ for the Euclidean probit model, $\ell\left(\mathbf{\Theta}\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{y_{i,j}\log\theta_{i,j} + \left(1 - y_{i,j}\right)\log(1 - \theta_{i,j})\right\}$, and the expectations and variances are computed with respect to the posterior distribution of the parameters (which are in turn approximated from the samples generated by our Markov chain Monte Carlo algorithm). The first term in the DIC can be understood as a goodness-of-fit measure, while the second one can be interpreted as a measure of model complexity. Note that, in every case, DIC favors the circular model.

2.5 Discussion

Our results suggest that the circular voting model developed in this chapter provides a better explanation for voting patterns in the modern U.S. House of Representatives than traditional Euclidean models. This increasing circularity, driven by the raise of extreme ideological factions willing to vote against the mainstream of the party (especially among Republicans) seems to have gone hand in hand with increasing polarization in the chamber.

In our interactions with with various colleagues we have heard two potential criticisms of the approach discussed in this section. The first one relates to whether the circular assumption makes mechanistic sense and, in particular, whether the Horseshoe Theory can be applied in the context of the U.S. House of Representatives. We view this criticism as a reflection of the type deductive thinking that tends to dominate in political science. In contrast, this chapter posits an inductive approach to the problem in which we use data to evaluate the empirical support for the Horseshoe Theory. The main contribution of this chapter is to provide

a solid statistical methodology that enables this kind empirical explorations, under the much laxer assumption that a circular voting space is at least minimally plausible. We recognize that taking an inductive approach begs the question of whether there are alternative explanations (e.g., heavy-tailed random shocks to the latent utilities or non-monotonic utilities functions) for the "extremes voting together" phenomenon. We agree that further exploration of these questions is key, but also see such endeavors as beyond the scope of this particular document, which is focused on introducing the basic methodology needed to fit and assess evidence related to models that rely on circular policy spaces. Furthermore, the fact that our results partially challenge the *ex-ante* opinion that the Horseshoe Theory might not be a good fit in U.S. that some experts might hold should be seen as a finding worth further investigation rather than a reason to discard the underlying methodology enabling those conclusion.

A second criticism relates to whether the additional complexity introduced by mapping the ideal points onto a circular space is needed when we could simply fit a two-dimensional Euclidean model. This criticism is often leveled even after conceding that the policy space might indeed be circular. One simple answer is *parsimony*. It is true that the circle is a sub-manifold of $\mathbb{R}^2$, but it is one with a lower intrinsic dimension (recall the discussion in Section 2.2.4). As a consequence, the number of parameters required to fit our circular model is smaller than the number of parameters in a 2D Euclidean model. A basic tenet in science is that among models that provide similar fit, the most parsimonious one should be preferred. More importantly, this criticism ignores one of the key messages of this chapter, that the dimensionality and the geometry of the policy space are related but subtly different concepts. This is a point that [47] made almost 50 years ago but is still not fully appreciated.

The focus of this chapter has been on circular models, which are essentially unidimensional. In the next chapter, we extend it to situations in which the latent space corresponds to high-dimensional spheres. While the likelihood formulation is straightforward, a key challenge in this setting is the elicitation of prior distributions that behave appropriately as the number of dimensions increase.

Table 2.4: DIC for the the circular and Euclidean model fitted to the 100^{th} to 116^{th} U.S. House of Representatives. Results for the 116^{th} House include only the first session.

House	I	J	Euclidean (1D)	Circular
100^{th}	425	939	-91829	**-89703**
101^{th}	429	879	-101183	**-99959**
102^{th}	431	901	-101021	**-99973**
103^{th}	430	1094	-112526	**-110386**
104^{th}	428	1321	-129709	**-127894**
105^{th}	426	1166	-112902	**-110475**
106^{th}	432	1209	-107545	**-103502**
107^{th}	428	990	-70693	**-66771**
108^{th}	430	1218	-72599	**-68973**
109^{th}	430	1210	-86097	**-82397**
110^{th}	424	1865	-96678	**-93227**
111^{th}	426	1647	-74824	**-70219**
112^{th}	428	1602	-118707	**-113963**
113^{th}	424	1202	-75144	**-71718**
114^{th}	431	1322	-67848	**-64067**
115^{th}	450	1207	-55772	**-51162**
116^{th}	432	700	-33948	**-30110**

Chapter 3

Spherical Factor Model for Binary Data

In the previous chapter, we proposed the univariate dimensional circular factor model which embeds binary data to a circular space by exploiting the random utility formulation underlying binary regression models. We also demonstrated that circular latent spaces offer a more consistent explanation for legislators's stated policy positions in the House of Representatives than Euclidean models. This chapter builds on ideas proposed in Chapter 2 to generate a general framework for embedding binary data into general spherical latent spaces, in a way that includes traditional factor models as a special case. This chapter focuses on two key challenges. The first one is eliciting prior distributions for model parameters that do not degenerate as the dimension of the embedding space grows. Well calibrated priors, in the previous sense, are key to enable dimensionality selection. The second challenge is designing efficient computational algorithms for a model in which some of the parameters live on a Riemannian manifold.

The remainder of the chapter is organized as follows: Section 3.1 introduces our spherical factor models and discusses the issues associated with prior elicitation. Section 3.2 discusses our computational approach to estimating model parameters, which is based on Geodesic Hamiltonian Monte Carlo algorithms. Section 3.3 illustrates the behavior of our model on both simulated and real datasets. Finally, Section 3.4 concludes the chapter with a discussion of future directions for our work.

3.1 Bayesian factor models for binary data on spherical spaces

3.1.1 Likelihood formulation

As in Sections 1.2 and 2.2, we also formulate our model in this chapter under the random utility framework. Consider a sequence of multivariate responses $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$, where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,J})$, and introduce parameters $\boldsymbol{\beta}_i$, $\boldsymbol{\psi}_j$ and $\boldsymbol{\zeta}_j$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ such that $\boldsymbol{\beta}_i, \boldsymbol{\psi}_j, \boldsymbol{\zeta}_j \in \mathcal{S}^K$, the unit hypersphere in $K+1$ dimensions, we can embed the positions $\boldsymbol{\beta}_i$, $\boldsymbol{\psi}_j$ and $\boldsymbol{\zeta}_j$ into a connected Riemannian manifold $\mathcal{D}$ equipped with a metric $\rho : \mathcal{D} \times \mathcal{D} \to M \subseteq \mathbb{R}^+$. In this chapter, we focus on the case in which $\mathcal{D}$ corresponds to the unit hypersphere in $K+1$ dimensions, $\mathcal{S}^K$, which is equipped with its geodesic distance ρ_K, the shortest angle separating two points measured over the great circle connecting them. Rewriting the utility function in 2.2, we obtain

$$U_+(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i) = -\left\{\rho_K\left(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i\right)\right\}^2 + \epsilon_{i,j}, \quad U_-(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i) = -\left\{\rho_K\left(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)\right\}^2 + \nu_{i,j}, \quad (3.1)$$

where the errors $\epsilon_{i,j}$ and $\nu_{i,j}$ are such that their differences $\upsilon_{i,j} = \nu_{i,j} - \epsilon_{i,j}$ are independent with cumulative distribution function G_j. This formulation leads to a likelihood function of the form

$$p\left(\boldsymbol{Y} \mid \{\boldsymbol{\psi}_j\}_{j=1}^{J}, \{\boldsymbol{\zeta}_j\}_{j=1}^{J}, \{\boldsymbol{\beta}_i\}_{i=1}^{I}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} \theta_{i,j}^{y_{i,j}} \left(1-\theta_{i,j}\right)^{1-y_{i,j}},$$

where $\boldsymbol{Y}$ is the $I \times J$ matrix of observations whose rows correspond to $\boldsymbol{y}_i^T$, and

$$\theta_{i,j} = G_j\left(e\left(\boldsymbol{\psi}_j, \boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)\right), \quad e\left(\boldsymbol{\psi}_j, \boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right) = \left\{\rho_K\left(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)\right\}^2 - \left\{\rho_K\left(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i\right)\right\}^2. \quad (3.2)$$

There are two alternative ways to describe $\mathcal{S}^K$. The first one uses a hyperspherical coordinate system and relies on a vector of K angles, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K) \in [-\pi, \pi] \times [-\pi/2, \pi/2]^{K-1}$. The second one uses the fact that $\mathcal{S}^K$ is a submanifold of $\mathbb{R}^{K+1}$, and relies on a vector of $K+1$ Cartesian coordinates $\boldsymbol{x} = (x_1, \ldots, x_{K+1})$ subject to the constraint $\|\boldsymbol{x}\| = 1$. The two representations are connected through the transformation

$$\begin{aligned}
x_1 &= \cos\phi_1 \cos\phi_2 \cos\phi_3 \cdots \cos\phi_{K-1}, \\
x_2 &= \sin\phi_1 \cos\phi_2 \cos\phi_3 \cdots \cos\phi_{K-1}, \\
x_3 &= \sin\phi_2 \cos\phi_3 \ldots \cos\phi_{K-1} \\
&\vdots \\
x_K &= \sin\phi_{K-1} \cos\phi_K, \\
x_{K+1} &= \sin\phi_K.
\end{aligned} \quad (3.3)$$

While the hyperspherical coordinate representation tends to be slightly more interpretable and directly reflects the true dimensionality of the embedding space, Cartesian coordinates often result in more compact expressions. For example the distance metric ρ_K can be simply written as $\rho_K(\boldsymbol{x}, \boldsymbol{z}) = \arccos\left(\boldsymbol{x}^T\boldsymbol{z}\right)$. Further-

more, the Cartesian coordinate representation will be key to the development of our computational approaches. Notation-wise, in the sequel, we adopt the convention of using Greek letters when representing points in $\mathcal{S}^K$ through their angular representation, and using Roman letters when representing them through embedded Cartesian coordinates.

Lastly, we discuss the selection of the link function. G_j must account for the fact that, when $\mathcal{S}^K$ is used as embedding space and ρ_K as the distance metric, the function $e\left(\boldsymbol{\psi}_j, \boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)$ introduced in Equation (3.2) has as its range the interval $[-\pi^2, \pi^2]$. Similar to the reasons outlined in 2.2.1, we again opt to work with the cumulative distribution of a shifted and scaled beta distribution,

$$G_j(z) = G_{\kappa_j}(z) = \int_{-\pi^2}^{z} \frac{1}{2\pi^2} \frac{\Gamma(2\kappa_j)}{\Gamma(\kappa_j)\Gamma(\kappa_j)} \left(\frac{\pi^2 + z}{2\pi^2}\right)^{\kappa_j - 1} \left(\frac{\pi^2 - z}{2\pi^2}\right)^{\kappa_j - 1} dz, \quad z \in [-\pi^2, \pi^2]. \tag{3.4}$$

As noted in Section 2.2.1, $1/\sqrt{\kappa_j}$, which controls the dispersion of the density g_j associated with G_j, plays an analogous role to the one that σ_j played in Section 1.2. Furthermore, recall that, for large values of κ_j the density of this Beta distribution is well approximated by a Gaussian distribution (please see Equation 2.4).

This observation, together with the fact that $\rho_K(\boldsymbol{x}, \boldsymbol{z}) = \arccos\left(\boldsymbol{x}^T \boldsymbol{z}\right) \approx \|\boldsymbol{x} - \boldsymbol{z}\|$ for small values of $\|\boldsymbol{x} - \boldsymbol{z}\|$, make it clear that the projection of the likelihood of a spherical model in $\mathcal{S}^K$ on its tangent bundle is very close to a version of the Euclidean factor model in $\mathbb{R}^K$ described in Equation (1.1) in which the link function is the probit link. We expand on this connection in Section 3.1.3

Finally, a short note on the connection between the likelihood functions associated with two models defined in $\mathcal{S}^K$ and $\mathcal{S}^{K-1}$, respectively. Let $\boldsymbol{\beta}, \boldsymbol{\psi} \in \mathcal{S}^K$ and

$\boldsymbol{\beta}^*, \boldsymbol{\psi}^* \in \mathcal{S}^{K-1}$ be (the angular representations of) two pairs of points in spherical spaces of dimension K and $K-1$, respectively. It is easy to verify that, when $\beta_K = \psi_K = 0$ as well as $\beta_k = \beta_k^*$ and $\psi_k = \psi_k^*$ for all $k = 1, \ldots, K-1$, then $\rho_K(\boldsymbol{\beta}, \boldsymbol{\psi}) = \rho_{K-1}(\boldsymbol{\beta}^*, \boldsymbol{\psi}^*)$. Hence, the likelihood function of a spherical factor model in which the latent positions are embedded in $\mathcal{S}^K$ and for which $\beta_{i,K} = \psi_{j,K} = \zeta_{j,K} = 0$ for all i and j, is identical to the likelihood function that would be obtained by directly fitting a model in which the latent positions are embedded in $\mathcal{S}^{K-1}$. This observation will play an important role in the next Section, which is focused on defining prior distributions for the latent positions.

3.1.2 Prior distributions

As we discussed in the beginning of this chapter, a key challenge involved in the implementation of the spherical factor models we just described is selecting priors for the latent positions $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_J$ and $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_J$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$. This challenge is specially daunting in settings where estimating the dimension K of the latent space is of interest.

To illustrate the issues involved, consider the use of the von Mises-Fisher family as priors for the latent positions. The Hausdorff density with respect to the uniform measure on the sphere for the von Mises-Fisher distribution on $\mathbb{R}^d$ is given by

$$p_{\mathcal{H}}(\boldsymbol{x} \mid \boldsymbol{\eta}, \omega) = \frac{\omega^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\omega)} \exp\left\{\omega \boldsymbol{\eta}^T \boldsymbol{x}\right\}, \qquad \|\boldsymbol{x}\| = 1,$$

where $I_\nu(\cdot)$ is the modified Bessel function of order ν, $\boldsymbol{\eta}$ satisfies $\|\boldsymbol{\eta}\| = 1$ and represents the mean direction, and ω is a scalar precision parameter. The von Mises-Fisher distribution is a natural prior in this setting because it is the spherical

analogue to the multivariate Gaussian distribution with diagonal, homoscedastic covariance matrix that is sometimes used as a prior for the latent positions in traditional factor models. Indeed, note that

$$\lim_{\omega \to \infty} \frac{\frac{\omega^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\omega)} \exp\left\{\omega \boldsymbol{\eta}^T \boldsymbol{x}\right\}}{\left(\frac{\omega}{2\pi}\right)^{d/2} \exp\left\{-\frac{\omega}{2}(\boldsymbol{x}-\boldsymbol{\eta})^T(\boldsymbol{x}-\boldsymbol{\eta})\right\}} = 1.$$

We are interested in the behavior of the prior on

$$\theta_{i,j} = G_{\kappa_j}\left(\left\{\rho_K\left(\boldsymbol{\zeta}_j, \boldsymbol{\beta}_i\right)\right\}^2 - \left\{\rho_K\left(\boldsymbol{\psi}_j, \boldsymbol{\beta}_i\right)\right\}^2\right)$$

(recall Equation (3.2)) induced by a von Mises-Fisher prior with mean direction $\mathbf{0}$ and precision ω for $\boldsymbol{\beta}_i$, and independent von Mises-Fisher priors with mean direction $\mathbf{0}$ and precision τ for both $\boldsymbol{\psi}_j$ and $\boldsymbol{\zeta}_j$. Because, $\boldsymbol{\psi}_j$ and $\boldsymbol{\zeta}_j$ are assigned the same prior distributions, it is clear from a simple symmetry argument that $\mathrm{E}\left(\theta_{i,j}\right) = 1/2$ for all values of ω, τ and κ_j. On the other hand, Figure 3.1 shows the value $\mathrm{Var}\left(\theta_{i,j}\right)$ as a function of the embedding dimension K for various combinations of ω, τ and κ_j. Note that, in every case, it is apparent that $\lim_{K\to\infty} \mathrm{Var}\left(\theta_{i,j}\right) \to 0$, which implies that $\theta_{i,j}$ converges in distribution to a point mass at $1/2$ as the number of latent dimensions grows.

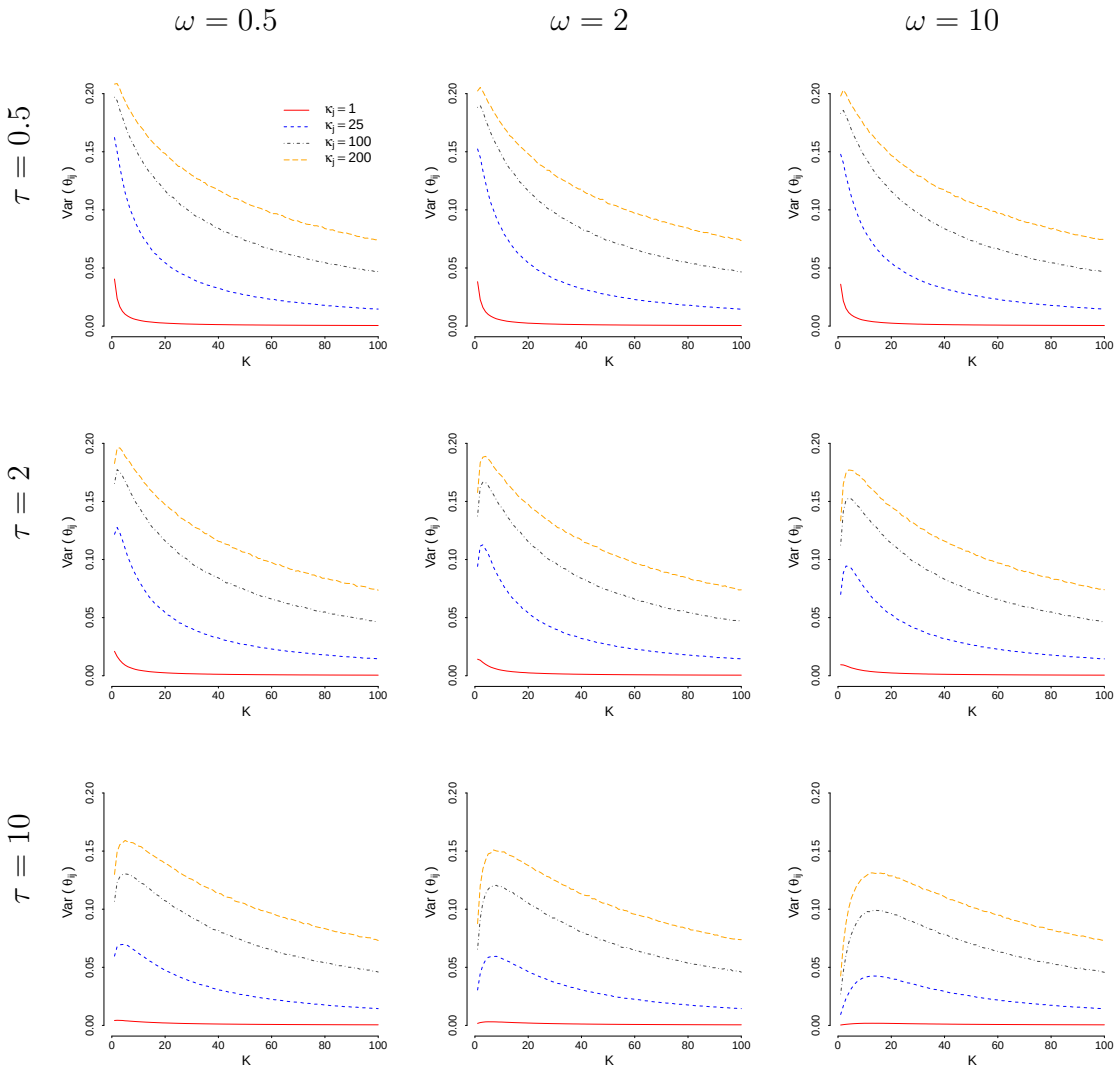


Figure 3.1: Prior variance of $\theta_{i,j}$ induced by von Mises-Fisher priors on the latent coordinates.

This behavior, which is clearly unappealing, is driven by two features. First, the von Mises-Fisher prior has a single scale parameter for all its dimensions. Secondly, the surface area of the K dimensional sphere, $\frac{2\pi^{K/2}}{\Gamma(K/2)}$, tends to zero as $K \to \infty$. In fact, not only the von Mises-Fisher distribution, but also other widely used distributions on the sphere such as the Fisher-Bingham and the Watson distributions

suffer from the same drawback. [83] proposes to overcome this phenomenon by scaling the (common) concentration parameter of the von Mises-Fisher according to the number of dimensions. However, this approach is unappealing in our setting for two reasons. First, it does not appear to overcome the degeneracy issues just discussed. Secondly, under this approach changes in K would fundamentally change the nature and structure of the prior distribution. In the sequel, we discuss a novel class of flexible priors distributions on $\mathcal{S}^K$ that allow for different scales across various dimensions and can be used to construct priors for $\theta_{i,j}$ that do not concentrate on a point mass as the number of dimensions increase.

We build our new class of prior distributions in terms of the vector of angles $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K) \in [-\pi, \pi] \times [-\pi/2, \pi/2]^{K-1}$ associated with the hyperspherical coordinate system. In particular, we assign each component of $\boldsymbol{\phi}$ a (scaled) von Mises distribution centered at 0, so that

$$p(\boldsymbol{\phi} \mid \boldsymbol{\omega}) = \left(\frac{1}{2\pi}\right)^K 2^{K-1} \frac{1}{I_0(\omega_1)} \exp\{\omega_1 \cos\phi_1\} \prod_{k=2}^{K} \frac{1}{I_0(\omega_k)} \exp\{\omega_k \cos 2\phi_k\}, \quad (3.5)$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$ is a vector of dimension-specific precisions. We call this prior the spherical von Mises distribution, and denote it by $\boldsymbol{\phi} \sim \text{SvM}(\boldsymbol{\omega})$.

One key feature of this prior is that the parameters $\omega_1, \ldots, \omega_K$ control the concentration of each marginal distribution around their mean. In particular, when $\omega_K \to \infty$, the marginal distribution of ϕ_K becomes a point mass at zero, and $p(\boldsymbol{\phi} \mid \boldsymbol{\omega})$ concentrates all its mass in a greater nested hypersphere of dimension $K - 1$ (see Figure 3.2 for an illustration).

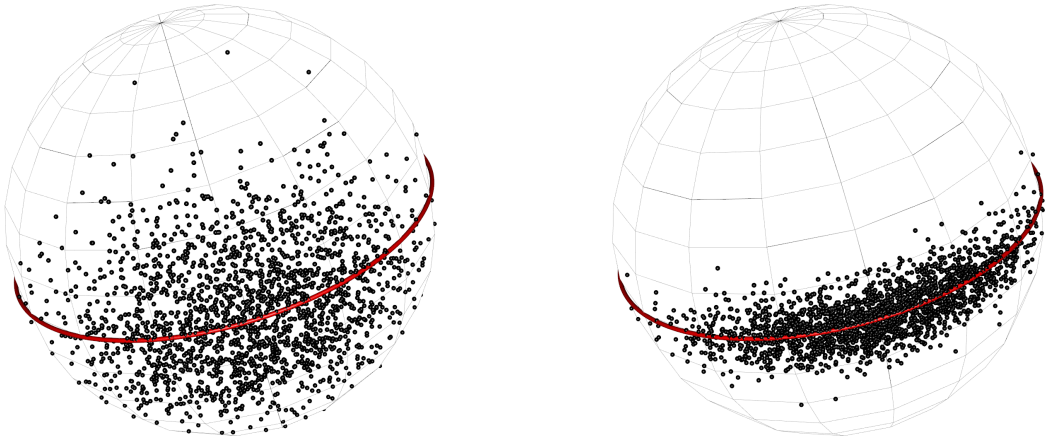


Figure 3.2: Draws from two spherical von Mises distributions in $\mathcal{S}^3$. The left panel corresponds to $\omega_1 = 7$ and $\omega_2 = 4$, while the right panel corresponds to $\omega_1 = 7$ and $\omega_2 = 30$. Note that, as ω_2 increases, the draws concentrate around a great circle.

A second key feature of this class of priors is that

$$\lim_{\boldsymbol{\omega}\to\infty} \frac{\left(\frac{1}{2\pi}\right)^K 2^{K-1} \frac{1}{I_0(\omega_1)} \exp\left\{\omega_1 \cos\phi_1\right\} \prod_{k=2}^K \frac{1}{I_0(\omega_k)} \exp\left\{\omega_k \cos 2\phi_k\right\}}{\left(\frac{1}{2\pi}\right)^{K/2} 2^{K-1} \left\{\prod_{k=1}^K \omega_k\right\}^{1/2} \exp\left\{-\frac{1}{2}\left[\omega_1\phi_1^2 + 4\sum_{k=2}^K \omega_k\phi_k^2\right]\right\}} = 1. \quad (3.6)$$

Somewhat informally, this means that for large values of the precision parameters, the prior behaves like a zero-mean multivariate normal distribution with covariance matrix $\text{diag}\{1/\omega_1, 1/(4\omega_2), \ldots, 1/(4\omega_{K-1})\}$.

Finally, we note that the density of the spherical von Mises distribution can be written in terms of the Cartesian coordinates $\boldsymbol{x} = (x_1, \ldots, x_{K+1})^T$ associated with $\boldsymbol{\phi}$ (recall Equation (3.3)). More specifically, the density of the Hausdorff measure with respect to the uniform distribution on the sphere associated with Equation

(3.5) is given by

$$p(\boldsymbol{x} \mid \boldsymbol{\omega}) = \frac{1}{\prod_{k=1}^{K} \sqrt{\sum_{t=1}^{k+1} x_t^2}} \frac{1}{2\pi I_o(\omega_1)} \exp\left\{\omega_1 \frac{x_1}{\sqrt{x_1^2 + x_2^2}}\right\} \left\{\prod_{k=2}^{K} \frac{1}{\pi I_o(\omega_k)}\right\} \exp\left\{-\sum_{k=2}^{K} \omega_k \left(2\frac{x_{k+1}^2}{\sum_{t=1}^{k+1} x_t^2} - 1\right)\right\}, \qquad \boldsymbol{x}^T\boldsymbol{x} = 1. \quad (3.7)$$

See Appendix A.3 for the derivation.

Coming back to our spherical factor model, we use the spherical von Mises distribution as priors for the $\boldsymbol{\psi}_j$s, $\boldsymbol{\zeta}_j$s and $\boldsymbol{\beta}_i$s. In particular, we set

$$\boldsymbol{\psi}_i \sim \text{SvM}(\tau, 2^2\tau, 3^2\tau, \ldots, K^2\tau) \quad (3.8)$$

$$\boldsymbol{\zeta}_j \sim \text{SvM}(\tau, 2^2\tau, 3^2\tau, \ldots, K^2\tau) \quad (3.9)$$

$$\boldsymbol{\beta}_j \sim \text{SvM}(\omega, 2^2\omega, 3^2\omega, \ldots, K^2\omega) \quad (3.10)$$

independently across all $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

By setting up the priors on the latent positions to have increasing marginal precisions, we can avoid the degeneracy issues that arose in the case of the von Mises-Fisher priors. This is demonstrated in Figure 3.3, which shows the value of $\text{Var}(\theta_{i,j})$ as function of the latent dimension K for various combinations of ω, τ and κ_j under (3.8-3.10). Note that, in every case, the variance of the prior decreases slightly with the addition of the first few dimensions, but then seems to quickly stabilize around a constant that is determined by the value of the hyperparameters. Figure 3.4 explores in more detail the shape of the implied prior distribution on $\theta_{i,j}$, as well as the effect of various hyperparameters. Note that the implied prior can either be unimodal (which typically happens for relatively large values ω and τ), or trimodal (which happens for low values of ω and τ).

The intuition behind the formulation in Equations (3.8-3.10) comes from the fact that these polynomial rates ensures that $\mathrm{Var}\left(\theta_{i,j}\right)$ converges to a constant with the number of dimensions under the Gaussian approximation in Equation (3.6).

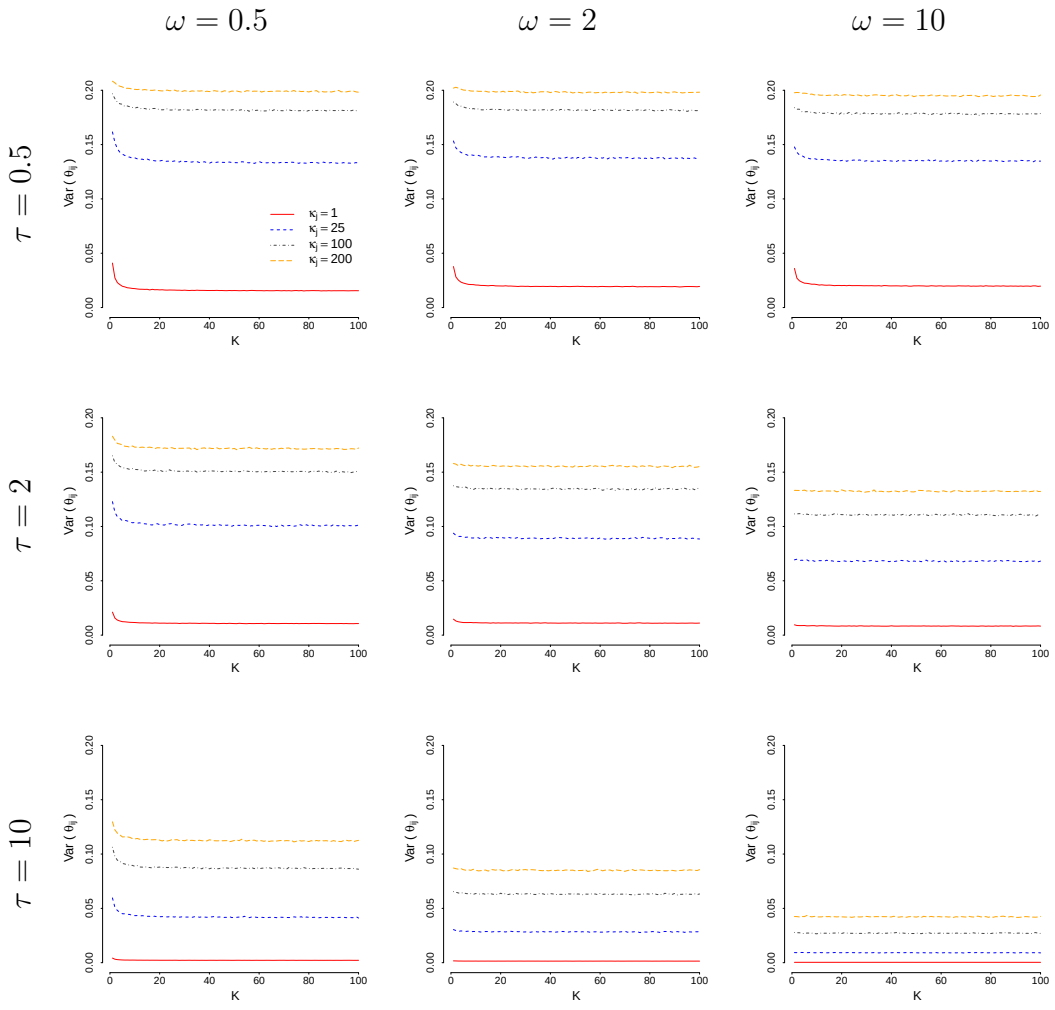
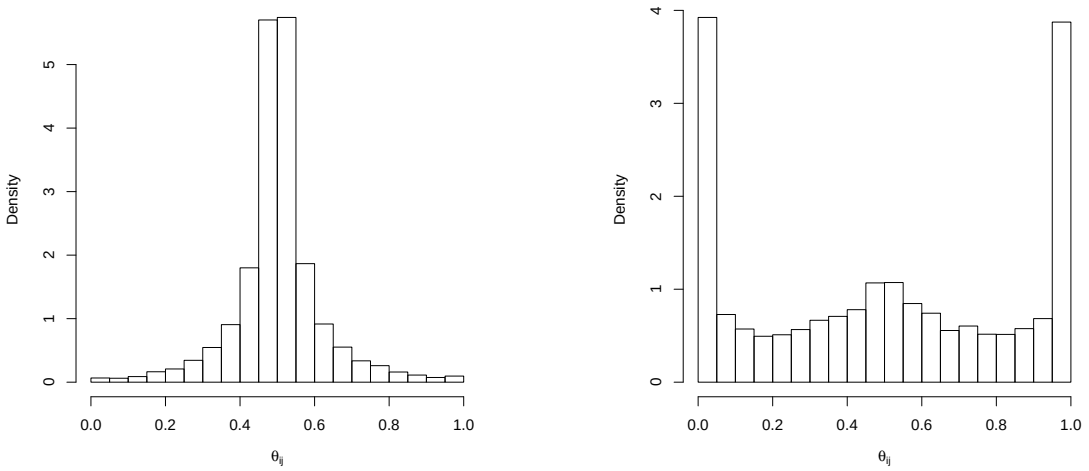


Figure 3.3: Prior variance of $\theta_{i,j}$ induced by spherical von Mises priors with polynomially increasing marginal precision on the latent coordinates.

In addition to avoiding degeneracy issues, using an increasing sequence for the marginal precision of all three sets of $\boldsymbol{\psi}_j$s, $\boldsymbol{\zeta}_j$s and $\boldsymbol{\beta}_i$s means that the prior encourages the model to concentrate most of the mass on a embedded sub-sphere

of $\mathcal{S}^K$ (recall the discussion at the end of Section 3.1.1). Hence, we can think of K as representing the highest intrinsic dimension allowed by our prior, and as the combination of ω and τ as controlling the "effective" prior dimension of the space, $K^* \leq K$. Another implication of this prior structure is that we can think of our prior as encouraging a decomposition of the (spherical) variance that is reminiscent of the principal nested sphere analysis introduced in [17].



(a) $\kappa_j = 200$, $\omega = 20$, $\tau = 20$ **(b)** $\kappa_j = 50$, $\omega = 1$, $\tau = 1$

Figure 3.4: Histogram of 10,000 draws from the prior distribution on $\theta_{i,j}$ implied by (3.8-3.10) for $K = 10$ and two different combinations of hyperparameters.

3.1.3 Connection to the Euclidean model

In Section 3.1.1 we argued that the spherical model of dimension K contains all other spherical models of lower dimension as special cases. Similarly, the Euclidean space of dimension $K + 1$ includes all spherical spaces of dimension K or lower. It is also true that the K-dimensional Euclidean model discussed in Section 1.2 (with a probit link) can be seen as a limit case of our spherical model on $\mathcal{S}^K$.

The argument, which is analogous to Section 2.2.4, relies on projecting the spher-

ical model onto the tangent bundle of the manifold, and making $\omega \to \infty$, $\tau \to \infty$ and $\kappa_j \to \infty$ while keeping ω/τ and ω/κ_j constant for all j. As mentioned in previous sections, under these circumstances, (i) the link function G_{κ_j} defined in Equation (3.4) projects onto the probit link function, (ii) the geodesic distance between the original latent positions converges to the Euclidean distance between their projections onto the tangent space, and (iii) the spherical von-Mises priors defined in Equations (3.8-3.10) project onto Gaussian priors.

This observation is important for at least three reasons. Firstly, it can guide prior elicitation, specially for the precision parameters ω and τ. Secondly, spherical models can be seen as interpolating (in terms of complexity) between Euclidean models of adjacent dimensions. In particular, note that the likelihood for Euclidean models involves $\mathcal{O}(\{I+2J\}K)$ parameters (since the σ_js are not identifiable and typically fixed) while the likelihood of the spherical model relies on $\mathcal{O}(\{I+2J\}K+J)$ parameters, and finally, $\mathcal{O}(\{I+2J\}K) < \mathcal{O}(\{I+2J\}K+J) < \mathcal{O}(\{I+2J\}\{K+1\})$. Thirdly, it enhances the interpretability of the model. In particular, the reciprocal of the precision, $\frac{1}{\omega}$ can be interpreted as a measure of sphericity of the latent space that can be directly contrasted across datasets.

3.1.4 Identifiability

The likelihood function for the spherical factor model discussed in Section 3.1.1 is invariant to simultaneous rotations of all latent positions. However, the structure of the priors in (3.8-3.10) induces weak identifiability for the $\boldsymbol{\psi}_j$s, $\boldsymbol{\zeta}_j$s, and $\boldsymbol{\beta}_i$s in the posterior distribution. Reflections, which are not addressed by our prior formulation, can be accounted for by fixing the octant of the hypersphere to which a particular $\boldsymbol{\beta}_i$ belongs. In our implementation, this constraint is enforced

a posteriori by post-processing the samples generated by the Markov chain Monte Carlo algorithm described in Section 3.2. On the other hand, while the scale parameters $\sigma_1, \ldots, \sigma_J$ are not identifiable in the Euclidean model, the analogous precisions $\kappa_1, \ldots, \kappa_J$ are identifiable in the spherical model. This is because $\mathcal{S}^K$ has finite measure, and therefore the likelihood is not invariant to rescalings of the latent positions. More generally, partial Procrustes analysis (which accounts for invariance to translations and rotations, but retains the scale) can be used to postprocess the samples to ensure identifiability.

3.1.5 Hyperpriors

Completing the specification of the model requires that we assign hyperpriors to ω, τ and $\kappa_1, \ldots, \kappa_J$. The precisions ω and τ are assigned independent Gamma distributions, $\omega \sim \text{Gam}(a_\omega, b_\omega)$ and $\tau \sim \text{Gam}(a_\tau, b_\tau)$. Similarly, the concentration parameters associated with the link function are assumed to be conditionally independent and given a common prior, $\kappa_j \sim \text{Gam}(c, \lambda)$, where λ is in turn given a conditionally conjugate Gamma hyperprior, $\lambda \sim \text{Gam}(a_\lambda, b_\lambda)$. The parameters a_ω, b_ω, a_τ, b_τ, a_λ, b_λ and c for these hyperpriors are assigned to strongly favor configurations in which $\beta_{i,1} \in [-\pi/2, \pi/2]$ (which is consistent with the assumption that a Euclidean model is approximately correct), and so that the induced prior distribution on $\theta_{i,j}$ is, for large K, close to a $\text{Beta}(1/2, 1/2)$ distribution (the proper, reference prior for the probability of a Bernoulli distribution). Various combinations of hyperparameters satisfy these requirements, most of which lead to priors on $\theta_{i,j}$ that are trimodal. We recommend picking one in which the hyperpriors are not very concentrated (e.g., see Figure 3.5) and studying the sensitivity of the analysis to the prior choice. In our experience, inferences tend

to be robust to reasonable changes in the hyperparameters (see Section 3.3.2).

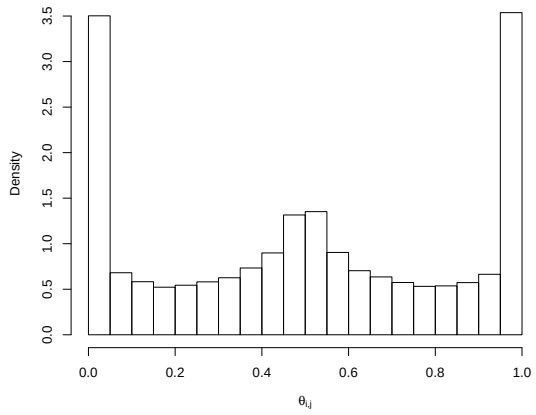


Figure 3.5: Histogram of 10,000 samples from the prior on $\theta_{i,j}$ implied by the hyperparameters $a_\omega = a_\tau = 1$, $b_\omega = 1/10$, $b_\tau = 5$, $a_\lambda = 2$, $b_\lambda = 150$, and $c = 1$ for $K = 10$. Unless otherwise noted, these are the parameter settings we use to carry out all the data analyses in this chapter.

3.2 Computation

The posterior distribution for the spherical factor model is analytically intractable. Hence, inference for the model parameters is carried out using Markov chain Monte Carlo (MCMC) algorithms. The algorithm we propose, which is very similar to the one we introduced in Section 2.3, is a hybrid that combines Gibbs sampling, random walk Metropolis-Hastings and Hamiltonian Monte Carlo steps to generate samples from the full conditional distributions of each parameter. The simplest steps correspond to sampling the parameters ω, τ, λ and $\kappa_1, \ldots, \kappa_J$. More specifically, we sample λ from its inverse-Gamma full conditional posterior distribution, and sample ω, τ and each of the κ_js using random walk Metropolis Hastings with log-Gaussian proposals. The variance of the proposals for these steps are tuned so that the acceptance rate is roughly 40%. On the other hand, for

sampling the latent positions we again employ the Geodesic Hamiltonian Monte Carlo (GHMC) algorithm described in Section 1.3.3.

As an example, consider the step associated with updating $\boldsymbol{\beta}_i$, the latent factor for individual i. Denoting the associated coordinates in $\mathbb{R}^{K+1}$ by $\boldsymbol{x}_{\beta_i}$ (recall Equation (3.3)), the density of the Hausdorff measure associated with the full conditional posterior is given by

$$p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots) \propto \left[\prod_{j=1}^{J} G_{\kappa_j}(e_{ij})^{y_{ij}} \left(1 - G_{\kappa_j}(e_{ij})\right)^{1-y_{ij}}\right] \left[\exp\left\{\omega \frac{x_{\beta_i,1}}{\sqrt{x_{\beta_i,1}^2 + x_{\beta_i,2}^2}}\right\}\right] \left[\exp\left\{-\sum_{k=2}^{K} k^2 \omega \left(2\frac{x_{\beta_i,k+1}^2}{\sum_{t=1}^{k+1} x_{\beta_i,t}^2} - 1\right)\right\}\right] \left[\frac{1}{\prod_{k=1}^{K}\left(\sum_{t=1}^{k+1} x_{\beta_i,t}^2\right)^{\frac{1}{2}}}\right], \qquad \boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1,$$

where $\boldsymbol{x}_{\beta_i} = (x_{\beta_i,1}, x_{\beta_i,2}, \ldots, x_{\beta_i,K+1})$. Then, given tuning parameters L and ϵ, the GHMC step takes the form:

1. Initialize $\boldsymbol{x}_{\beta_i} = \boldsymbol{x}_{\beta_i}^{(c)}$, as well as the auxiliary momentum variable $\boldsymbol{\gamma}$ by sampling $\boldsymbol{\gamma} \sim N(0, \boldsymbol{I}_{K+1})$.

2. Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\gamma} \leftarrow \left(\boldsymbol{I}_{K+1} - \boldsymbol{x}_{\beta_i}\boldsymbol{x}_{\beta_i}^T\right)\boldsymbol{\gamma}$, and then set $\boldsymbol{\gamma}^{(c)} = \boldsymbol{\gamma}$.

3. For each of the L leap steps:

   (a) Update the momentum by setting $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + \frac{\epsilon}{2}\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots)$.

   (b) Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\gamma} \leftarrow \left(\boldsymbol{I}_{K+1} - \boldsymbol{x}_{\beta_i}\boldsymbol{x}_{\beta_i}^T\right)\boldsymbol{\gamma}$, and then set the angular velocity of the geodesic flow $\nu = \|\phi\|$.

(c) Update $\boldsymbol{x}_{\beta_i}$ and $\boldsymbol{\gamma}$ jointly according to the geodesic flow with step size of ϵ,

$$\boldsymbol{x}_{\beta_i} \leftarrow \boldsymbol{x}_{\beta_i} \cos(\nu\epsilon) + \frac{\phi}{\nu} \sin(\nu\epsilon), \qquad \phi \leftarrow \phi \cos(\nu\epsilon) - \nu\, \boldsymbol{x}_{\beta_i} \sin(\nu\epsilon).$$

(d) Update $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + \frac{\epsilon}{2} \nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots)$.

(e) Project the momentum onto the tangent space at $\boldsymbol{x}_{\beta_i}$ by setting $\boldsymbol{\gamma} \leftarrow \left(\boldsymbol{I}_{K+1} - \boldsymbol{x}_{\beta_i} \boldsymbol{x}_{\beta_i}^T\right) \boldsymbol{\gamma}$.

4. Set the proposed values as $\boldsymbol{x}_{\beta_i}^{(p)} = \boldsymbol{x}_{\beta_i}$ and the proposed value $\boldsymbol{x}_{\beta_i}^{(p)}$ is accepted with probability

$$\min \left\{ 1, \frac{p_{\mathcal{H}} \left(\boldsymbol{x}_{\beta_i}^{(p)} \mid \cdots \right) \exp \left\{ -\frac{1}{2} \left[\boldsymbol{\gamma}^{(p)} \right]^T \boldsymbol{\gamma}^{(p)} \right\}}{p_{\mathcal{H}} \left(\boldsymbol{x}_{\beta_i}^{(c)} \mid \cdots \right) \exp \left\{ -\frac{1}{2} \left[\boldsymbol{\gamma}^{(c)} \right]^T \boldsymbol{\gamma}^{(c)} \right\}} \right\}$$

The gradient of the logarithm of the Hausdorff measure may appear forbidding to derive. One practical difficulty is dealing with the spherical constraint $\boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1$. We discuss the calculation of the gradient in Appendix A.4, where we also derive a recursive formula linking the gradient of the prior density in dimensions K to that of the gradient in dimension $K - 1$. Detailed expressions for the Hausdorff measures associated with the full conditional posteriors of the $\boldsymbol{x}_{\beta_i}$s, $\boldsymbol{x}_{\psi_j}$s and $\boldsymbol{x}_{\zeta_j}$s, as well as their corresponding gradients, can also be found Appendix A.4. In our experiments, we periodically "jitter" the step sizes and the number of leap steps (e.g., see 77, pg. 306). The specific range in which ϵ and L move for each (group of) parameter and each dataset is selected to target an average acceptance probability between 60% and 90% [65, 67].

3.3 Illustrations

In this section, we illustrate the performance of the proposed model on both simulated and real data sets. In all of these analyses, the number of leaps used in the HMC steps is randomly selected from a discrete uniform distribution between 1 and 10 every 50 samples. Similarly, the leap sizes are drawn from uniform distribution on $(0.01, 0.03)$ or $(0.01, 0.05)$ or $(0.005, 0.03)$ for each $\boldsymbol{\beta}_i$, and from a uniform distribution on $(0.01, 0.07)$ or $(0.01, 0.105)$ or $(0.01, 0.05)$ for each $\boldsymbol{\zeta}_j$ and $\boldsymbol{\psi}_j$. All inference presented in this Section are based on 20,000 samples obtained after convergence of the Markov chain Monte Carlo algorithm. The length of the burn in period varied between 10,000 and 20,000 iterations depending on the dataset, with a median around 10,000. Convergence was checked by monitoring the value of the log-likelihood function, both through visual inspection of the trace plot, and by comparing multiple chains using the procedure in [78].

3.3.1 Simulation study

We conducted a simulation study involving four distinct scenarios to evaluate our spherical model. For each scenario, we generate one data set consisting of $J = 700$ items and $I = 100$ subjects (similar in size to the roll call data from the U.S. Senate presented in Section 3.3.3). In our first three scenarios, the data is simulated from spherical factor models on $\mathcal{S}^2$, $\mathcal{S}^3$ and $\mathcal{S}^5$, respectively. In all cases, the subject-specific latent positions, as well as the item-specific latent positions, are sampled from spherical von-Mises distributions where all component-wise precisions are equal to 2. Note that this data generation mechanism is slightly different from the model we fit to the data (in which the precision of the components increase

with the index of the dimension). For the fourth scenario, data was simulated from a Euclidean probit model (recall Section 1.2) in which the intercepts $\mu_1, \ldots, \mu_J$, the discrimination parameters $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J$, and latent traits $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_I$ are sampled from standard Gaussian distributions.

In each scenario, both spherical and Euclidean probit factor models of varying dimensions are fitted to the simulated datasets. The left column of Figure 3.6 shows the value of the DIC as a function of the dimension K of the fitted model's latent space for each of the four scenarios in our simulation. Recall Section 2.4.3, we compute DIC for the Euclidean and spherical model similarly in this Chapter.

From Figure 3.6, we can see that DIC is capable of identifying the presence of sphericity in the underlying latent space, as well as its correct dimension. In Scenarios 1 to 3, the optimal model under DIC is correctly identified as a spherical model with the right dimension. Furthermore, as we would expect, the optimal Euclidean model in each case has one additional dimension (i.e., the dimension of the space corresponds to the lowest-dimensional Euclidean space in which the true spherical latent space can be embedded). For Scenario 4, DIC correctly selects a Euclidean model in three dimensions, followed very closely by a spherical space of the same dimension. Again, these results match our previous discussion about the ability of a spherical model to approximate a Euclidean model.

The second column of Figure 3.6 shows the in-sample predictive accuracy of the models in each scenario. This accuracy is closely linked to the goodness-of-fit component of the DIC. Note that, from the point of view of this metric, both the spherical and Euclidean models have about the same performance. Furthermore, in both cases, as the dimension increases, the predictive accuracy increases sharply at first, but quickly plateaus once we reach the right model dimension, generating

a clear elbow in the graph. Similar elbows can be seen in the third column of Figure 3.6, which shows the amount of variance associated with each component of a principal nested spheres decomposition [17] of the subject's latent positions for the $\mathcal{S}^{10}$. This decomposition can be interpreted as a version of principal component analysis for data on an spherical manifold. Unless stated otherwise, principal nested spheres decomposition is implemented with a fixed radius of 1 in this chapter. Note, however, that the elbow in Scenario 4 is much less sharp than the elbows we observed in Scenarios 1, 2 and 3.

To get a better sense of the relationship between the spherical and Euclidean models, Figure 3.7 presents histograms of the posterior samples for the hyperparameters ω, τ and $1/\lambda$ of our spherical model for Scenario 2 (left column) and Scenario 4 (right column). We focus on these two scenarios because the true dimension of the latent space is 3 in both cases. Note that the posterior distributions of these hyperparameters concentrate on very different values depending on whether the truth corresponds to a Euclidean or a spherical model. Furthermore, in all cases the posterior distributions are clearly different from the priors.

Finally, Figure 3.8 presents posterior estimates for the subject's latent positions for Scenario 1. To facilitate comparisons between the truth and the estimates, we plot each dimension separately. While the first dimension seems to be reconstructed quite accurately (the coverage rate for the 95% credible intervals shown in the left panel is 98%, with an approximate 95% confidence interval of $(0.953, 1.000)$), the reconstruction of the second component is less so (the coverage in this case is 68%, with an approximate 95% confidence interval of $(0.578, 0.782)$).

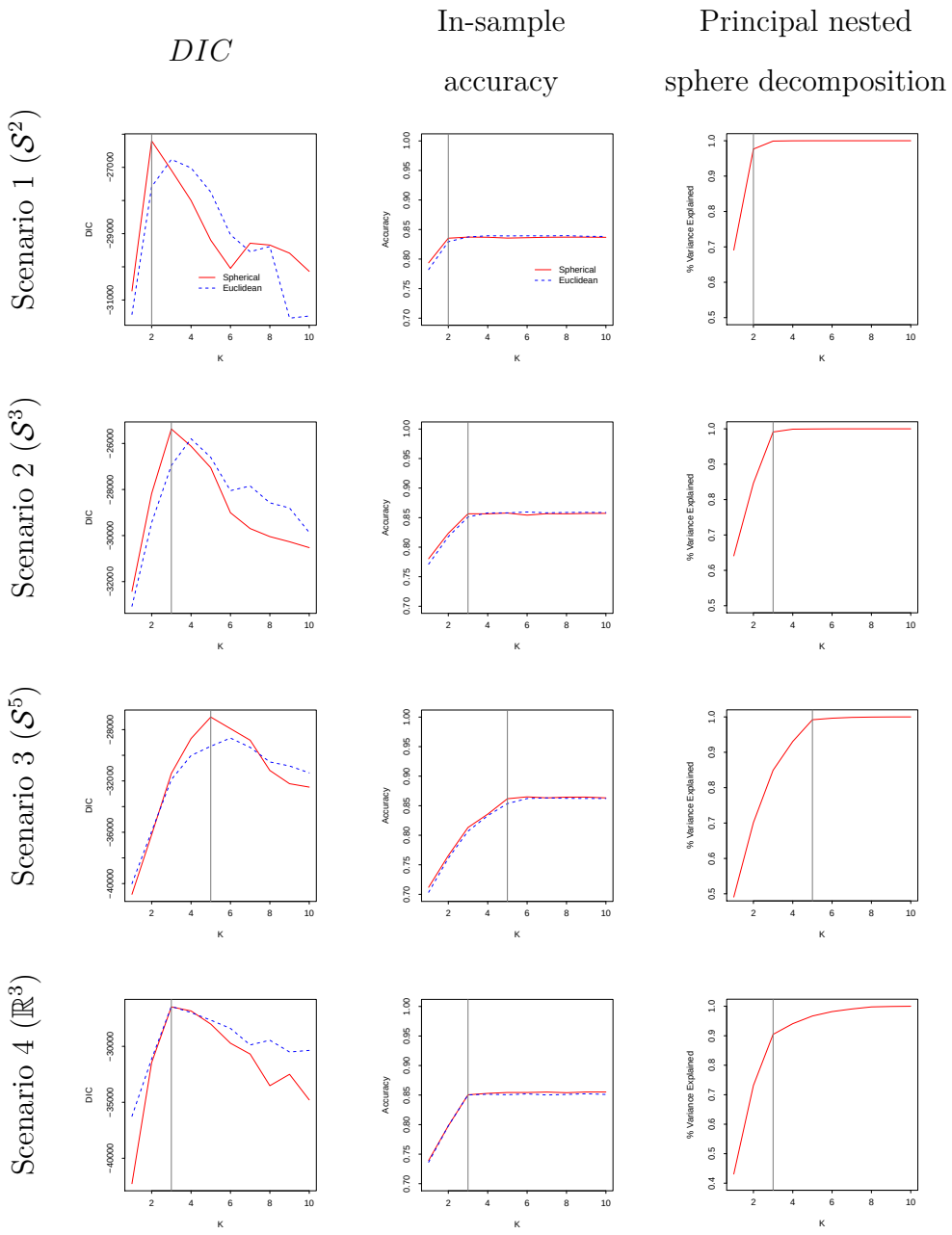


Figure 3.6: Deviance information criteria, in-sample predictive accuracy and principal nested sphere decomposition of the fitted models in each of the four simulation scenarios.

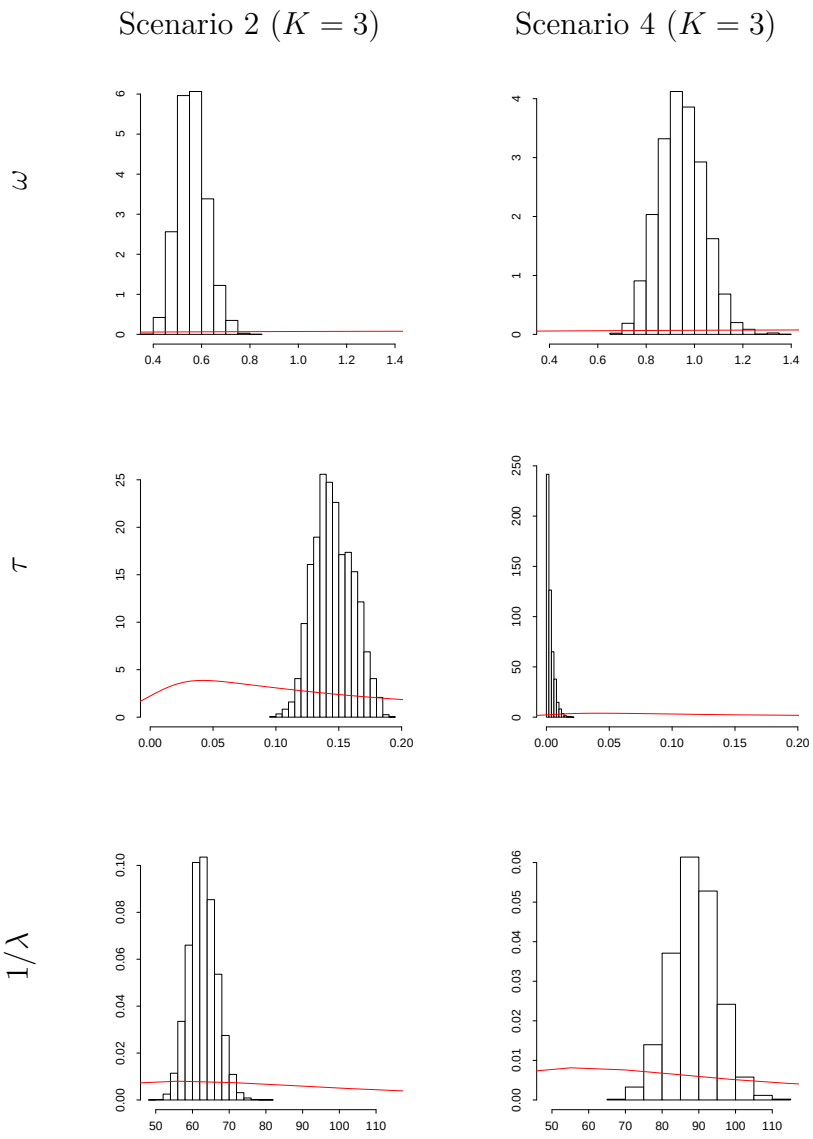
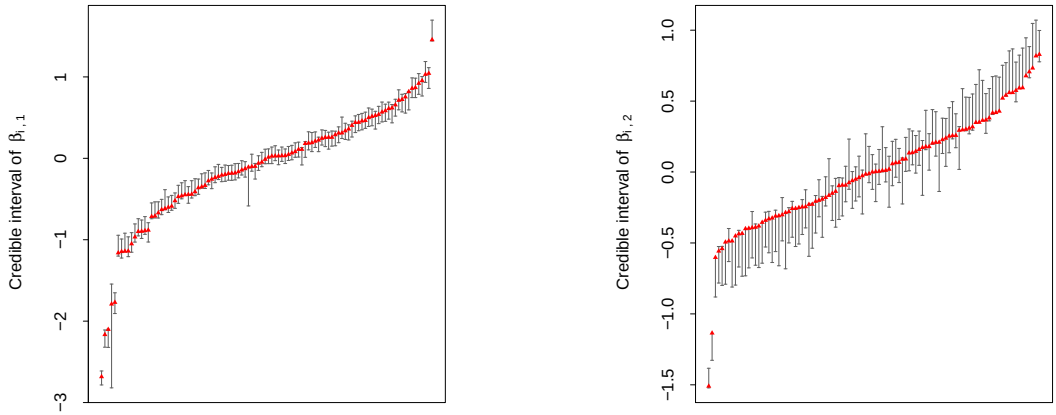


Figure 3.7: Histograms of the posterior samples for the hyperparameters ω, τ and $1/\lambda$ in scenarios 2 and 4 for $K = 3$. The continuous line corresponds to the prior distribution used in the analysis.



(a) Dimension 1 **(b)** Dimension 2

Figure 3.8: Scenario 1, 95% credible interval for the latent traits in each of the two dimensions. The true value of the traits is represented using a red triangle. Coverage rates are 99% and 68%, respectively.

3.3.2 Sensitivity analysis

To assess the sensitivity of our estimates to the values of the hyperparameters, we repeated our analysis using a different prior specification in which $\omega \sim \text{Gam}(1, 1/10)$, $\tau \sim \text{Gam}(1, 1/10)$ and $\lambda \sim \text{Gam}(2, 25)$. The induced prior on $\theta_{i,j}$ for $K = 10$ can be seen in Figure 3.9. (Note the very different shape when compared with Figure 3.5.)

We present here details only for Scenario 1; the results for the other 3 scenarios are very similar. Similarly to Figure 3.6, Figure 3.10 presents the value of the DIC, the in-sample accuracy and the principal nested sphere decomposition associated under both the original and the alternative priors. The main difference we observe is in the values of the DIC, with the alternative prior tending to give higher plausibility to models of dimension 5 and above.

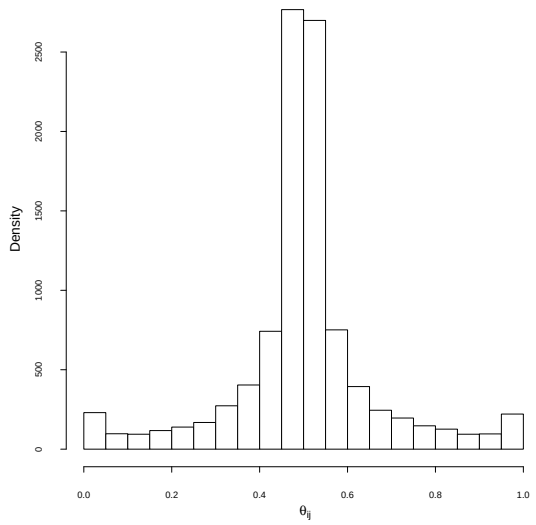
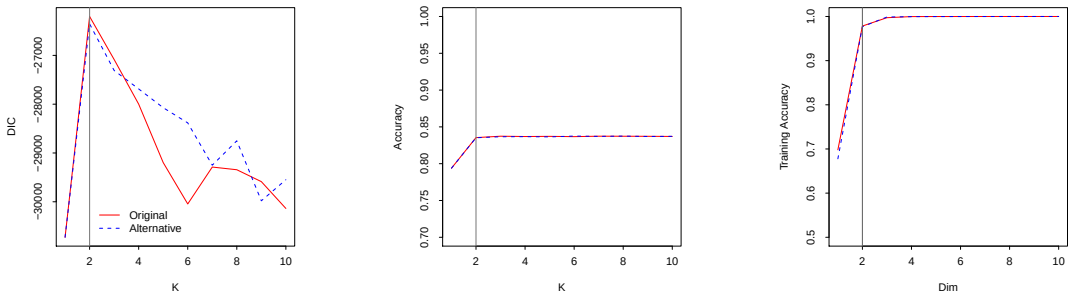


Figure 3.9: Histogram of 10,000 samples from the prior on $\theta_{i,j}$ implied by our alternative set of hyperparameters: $a_\omega = a_\tau = 1$, $b_\omega = b_\tau = 1/10$, $a_\lambda = 2$, $b_\lambda = 25$, and $c = 1$ for $K = 10$.



(a) *DIC* **(b)** In-sample accuracy **(c)** PNS decomposition

Figure 3.10: Deviance information criteria, in-sample predictive accuracy and principal nested sphere decomposition of the fitted models for Scenario 1 under our original and alternative priors.

Furthermore, Figure 3.11 shows histograms of the posterior distributions for the hyperparameters ω, τ and $1/\lambda$ under our two priors. The posterior distribution of these parameters appear to be very similar, suggesting robustness to this prior change. Finally, Figure 3.12 presents the posterior means of the subject's latent

positions under our alternative prior. Note that the point estimates are nearly identical under both priors.

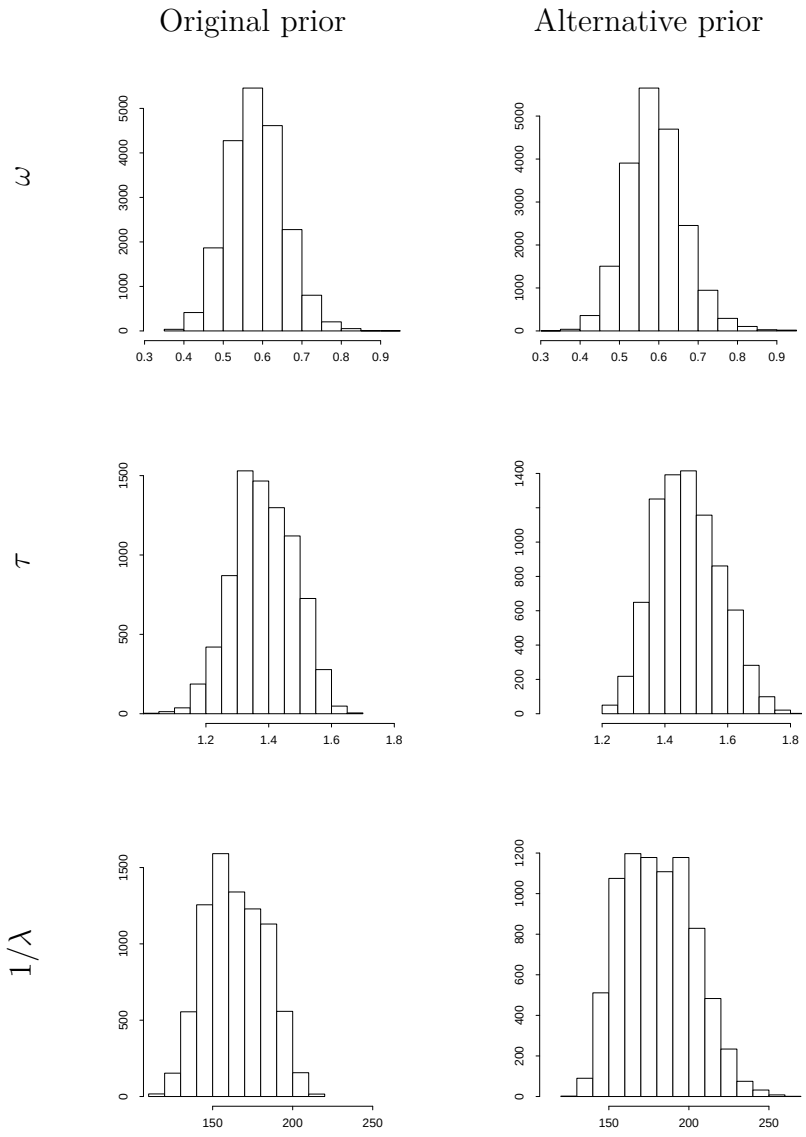
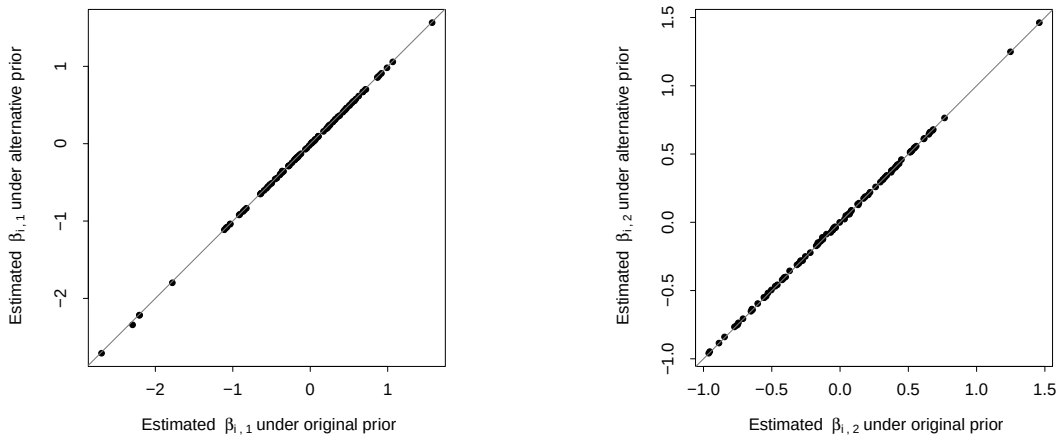


Figure 3.11: Histograms of the posterior samples for the hyperparameters ω, τ and $1/\lambda$ in Scenario 1 under the original (Figure 3.5) and the alternative (Figure 3.9) prior specifications.



(a) Dimension 1 **(b)** Dimension 2

Figure 3.12: Comparison of the posterior means of the locations $\boldsymbol{\beta}_i$ across two different prior for Scenario 1. Left panel compares the first dimension of the latent traits, while the right panel compares the values along the second dimension.

3.3.3 Roll call voting in the U.S. Senate

In this section we use our spherical factor models to investigate the geometry of policy spaces in two different U.S. Senates, the 102^{nd} (which met between January 3, 1991 and January 3, 1993, during the last two years of George H. W. Bush's presidency) and the 115^{th} (which met between January 3, 2017 and January 3, 2019, during the first two years of Donald Trump's presidency). We exclude from our analysis any senators that missed more than 40% of the votes cast during a given session, which substantially reduces the number missing values. The remaining ones, which are a small percentage of the total number of votes, were treated in our analysis as if missing completely at random. While this assumption is not fully supported by empirical evidence (e.g., see 52), it is extremely common in applications. Furthermore, the relatively low frequency of missing values in these datasets suggests that deviations from it will likely have a limited impact in

our results. See table 3.1 for a summary of the features of the two post processed datasets.

Session	Senators (I)	Measures (J)	Missing votes
102^{nd}	100	550	2222 (4.04%)
115^{th}	96	599	1236 (2.15%)

Table 3.1: Summary information for the two roll call datasets analyzed in this section.

As in Section 3.3.1, we fit both Euclidean and spherical factor models of varying dimensions to each of these two datasets. The left column of Figure 3.13 shows the DIC values associated with these models. For the 102^{nd} Senate, we can see that a Euclidean model of dimension 3 seems to provide the best fit overall while, among the spherical models, a model of dimension 3 also seems to outperform. On the other hand, for the 115^{th} Senate, a two-dimensional spherical model seems to provide the best fit to the data, closely followed by a circular (one-dimensional spherical) model. The right column of Figure 3.13 presents the principal nested sphere decomposition associated with the posterior mean configuration estimated by an eight-dimensional spherical model on each of the two Senates. The differences are substantial. In the case of the 115^{th} Senate (for which DIC suggests that the geometry of the latent space is indeed spherical), the first component of the decomposition explains more than 95% of the variability in the latent space, and the first three components explain close to 99%. On the other hand, for the 102^{nd} Senate, the first three spherical dimensions explain less than 70% of the total variability.

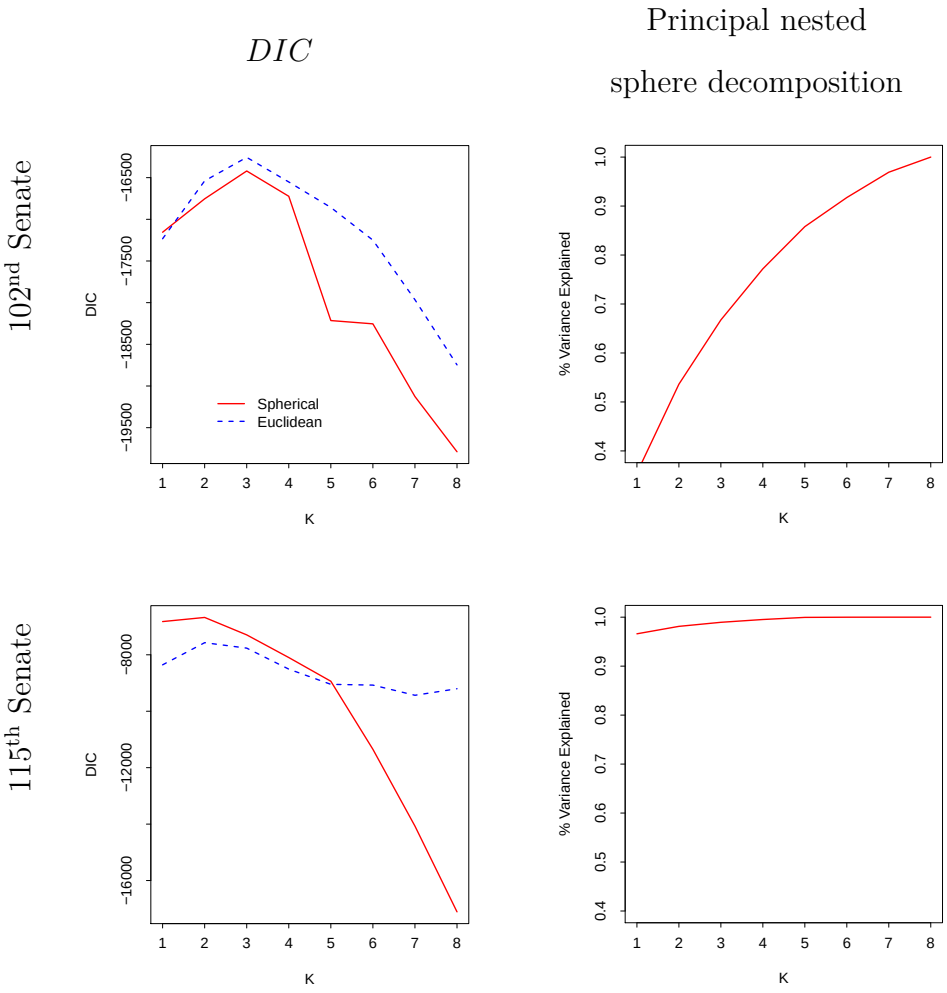


Figure 3.13: Deviance information criteria as a function of the embedding space's dimension K (left column) and principal nested sphere decomposition associated with the $\mathcal{S}^8$ model for the 102$^{\text{nd}}$ (top row) and the 115$^{\text{th}}$ (bottom row) U.S. Senates.

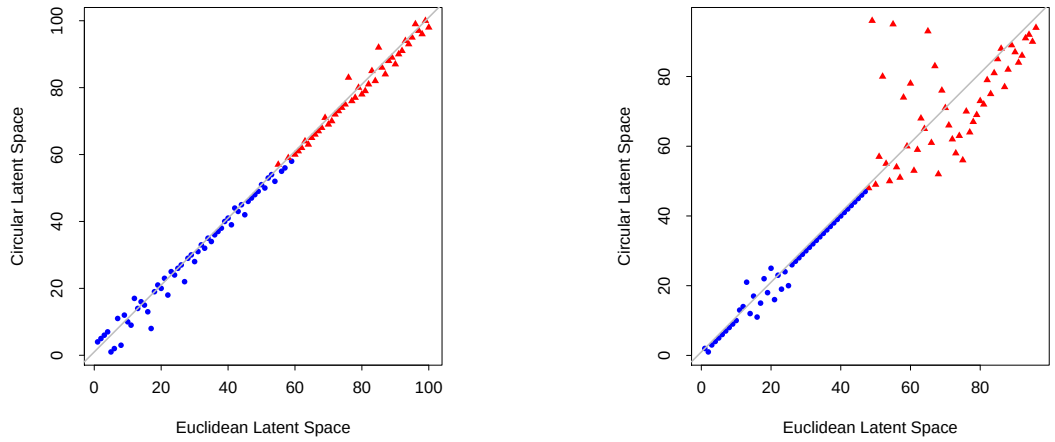
At a high level, these results are consistent with the understanding that scholars of American politics have of contemporaneous congressional voting patterns. While congressional voting in the U.S. has tended to be at least two-dimensional for most of its history (with one dimension roughly aligning with economic issues, and the other(s) corresponding to a combination of various social and cultural

issues), there is clear evidence that it has become more and more unidimensional over the last 40 years (e.g., see 84). Increasing polarization has also been well documented in the literature (e.g., see 25 and 85). The rise of extreme factions within both the Republican and Democratic parties willing to vote against their more mainstream colleagues, however, is a relatively new phenomenon that is just starting to be documented (see 79, 39, 40) and can explain the dominance of the spherical model for the 115$^{\text{th}}$ Senate voting data.

Figure 3.15 presents histograms of the posterior samples for the hyperparameters ω, τ and $1/\lambda$ for each of the two Senates. Note that, as with the simulations, the posterior distributions differ substantially from the priors. Furthermore, the model prefers relatively smaller values of ω and relatively larges values of τ and $1/\lambda$ for the 115$^{\text{th}}$ Senate when compared with the 102$^{\text{nd}}$.

To conclude this Section, we focus our attention on the one-dimensional versions of the Euclidean and spherical models. While one-dimensional models are not preferred by the DIC criteria in our datasets, researchers often use them in practice because the resulting ranking of legislators often reflects their ordering in a liberal-conservative (left-right) scale. Figure 3.14 compares the median rank order of legislators estimated under the one-dimensional Euclidean and one-dimensional spherical (circular) models for the 102$^{\text{nd}}$ (left panel) and the 115$^{\text{th}}$ (right panel) Senates. To generate the ranks under the circular model, we "unwrap" the circle and compute the ranks on the basis of the associated angles (which live in the interval $[-\pi, \pi]$). For the 102$^{\text{nd}}$ Senate (where DIC would prefer the one-dimensional Euclidean model over the circular model), the median ranks of legislators under both models are very similar. This result provides support to our argument that the spherical model can provide a very good approximation to the Euclidean model when there is no evidence of sphericity in the data. On the other hand, for the

115^{th} (where the circular model would be preferred over the one-dimensional Euclidean model), the ranking of Republican legislators differs substantially between both models. Digging a little bit deeper, we can see that the five legislators for which the ranks differ the most are Rand Paul (KY), Mike Lee (UT), Jeff Flake (AZ), Bob Corker (TN) and John Neely Kennedy (LA), all known for being hard line conservatives, but also for bucking their party and voting with (left wing) Democrats on some specific issues.



(a) 102^{nd} Senate **(b)** 115^{th} Senate

Figure 3.14: Comparison of the rank order of legislators between the 1D Euclidean and circular models for the 102^{nd} (left panel) and the 115^{th} (right panel) Senates. Red triangles correspond to Republican Senators, while blue circles indicate Democrats, as well as Independents who caucus with Democrats.

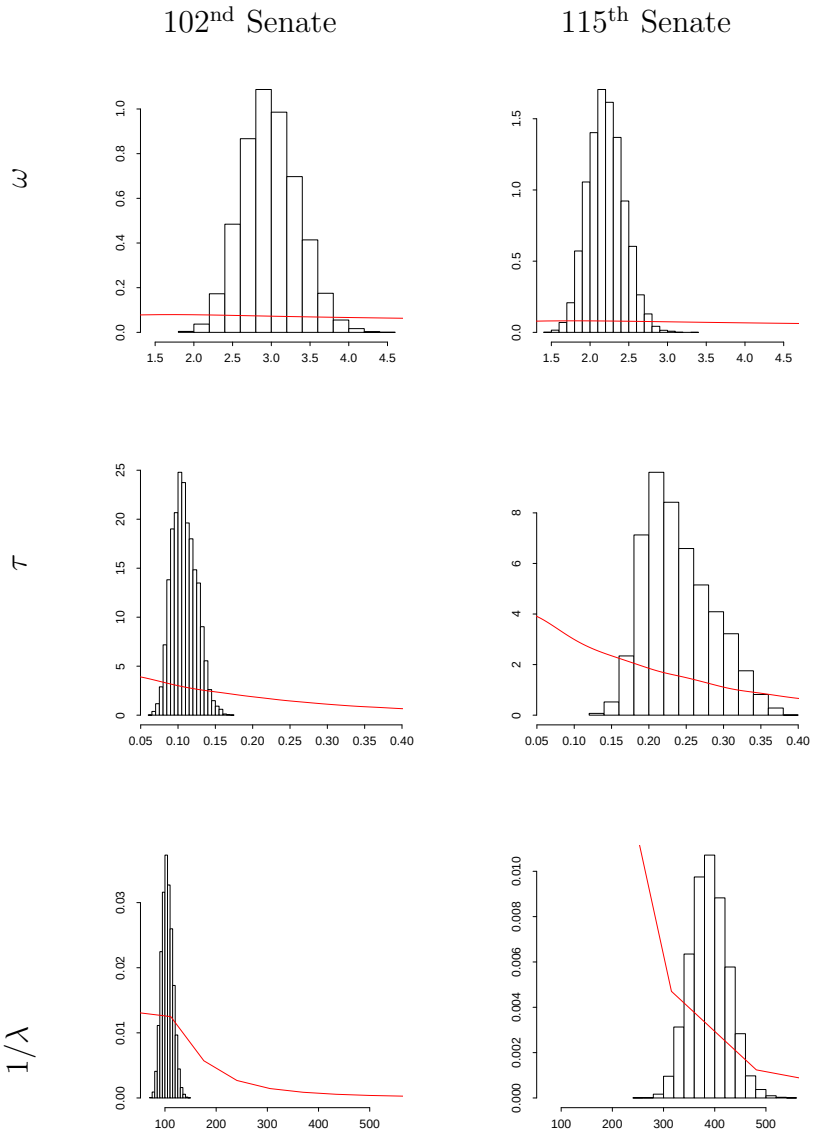


Figure 3.15: Histograms of the posterior samples for the hyperparameters ω, τ and $1/\lambda$ for the 102$^{\text{nd}}$ (left panel) and the 115$^{\text{th}}$ (right panel) U.S. Senates for $K = 3$ and $K = 2$, respectively. The continuous line corresponds to the prior distribution used in the analysis.

3.3.4 Roll call voting in the U.S. House of Representatives revisited

In this section, we revisit the two U.S. House of Representatives data sets analyzed in Section 2.4. Similar to Section 3.3.3, we fit both Euclidean and spherical factor models of varying dimensions to these two datasets. In the case of 112^{th} House, a spherical model of dimension 6 performs the best overall, followed closely by a spherical model of dimension 5. On the other hand, for the 116^{th} House, a spherical model of dimension 3 achieves the best performance overall, while a Euclidean of dimension 3 is the runner-up. In both cases, DIC favors the spherical latent space over the Euclidean one. As expected, the principal nested decompositions associated with the posterior mean configuration estimated by an nine-dimensional spherical model are quite different. In particular, the variance explained by the first three components for the 116^{th} House accounts for almost 99%. In contrary, the first three components only explain 75% of the variance.

3.4 Discussion

We have developed a new flexible class of factor analysis models for multivariate binary data that embeds the observations on spherical latent spaces. The results from our illustrations suggest that (i) it is possible to distinguish between different geometries and dimensions for the embedding space, (ii) our model can closely approximate traditional factor models when the latent space is Euclidean, (iii) the use spherical latent spaces can be justified, both theoretically and empirically, in applications related to choice models.

We consider two potential extensions as part of our future work. First we could consider ellipsoidal spaces, introduce a set of parameters that control the relative scale across dimensions. The second extension focuses on alternative priors that allow the ideal points to concentrate around more general sub manifolds of the $\mathcal{S}^K$. This second extension is motivated by some of our illustrations. In particular, panel (a) in Figure 3.17 shows the estimates of the ideal points for the legislators in the 116^{th} U.S. House of Representatives. Note that the points do not seem to concentrate around a great circle. Instead, they seem to roughly concentrate around a circle of radius less than one that does not align with either of the great circles around which our prior for the latent traits is centered. In other words, non-geodesic variation, which is discussed extensively in [17] seems to be present in this example. Hence, we plan to investigate the use of small sphere distributions such as Bingham-Mardia [86] and the small sphere distribution of the first and second kind [87] as a potential alternatives priors for the latent traits.

In the next chapter, we build on all ideas developed so far and extend our spherical models to also allow embeddings of ordinal data which is prevalent in survey applications.

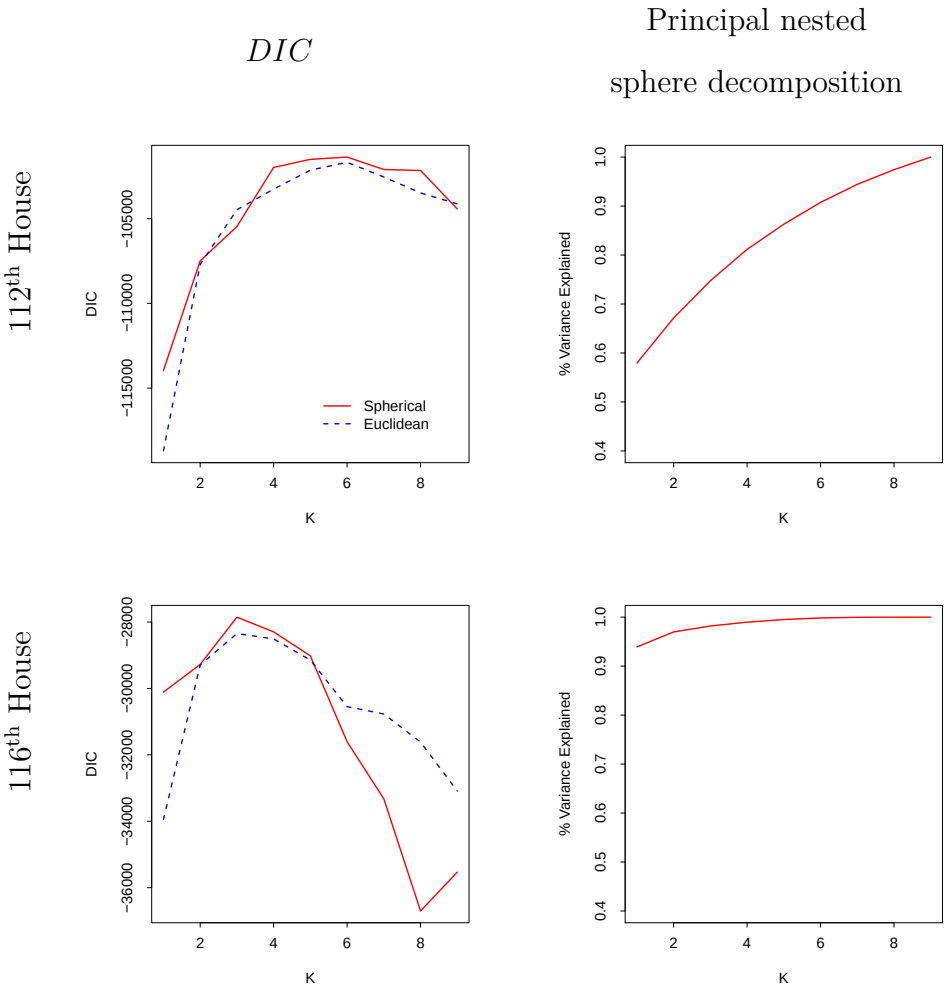
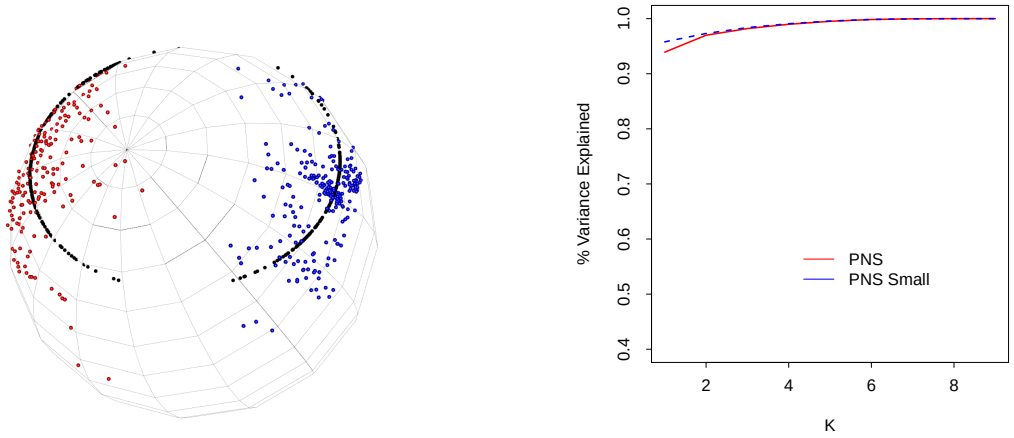


Figure 3.16: Deviance information criteria as a function of the embedding space's dimension K (left column) and principal nested sphere decomposition associated with the $\mathcal{S}^9$ model for the 112$^{\text{th}}$ (top row) and the 116$^{\text{th}}$ (bottom row) U.S. House of Representatives.



(a) PNS Small Decomposition

(b) Comparison of PNS and PNS Small Decomposition

Figure 3.17: Principal nested sphere decomposition associated with $\mathcal{S}^2$ model (left panel) and comparison of principal nested sphere decomposition and principal nested small sphere decomposition associated with $\mathcal{S}^9$ model for the 116^{th} U.S House of Representatives (right panel).

Chapter 4

Spherical Factor Model for Ordinal Data

In the previous two chapters, we developed a novel multivariate embedding method for binary data in which the underlying latent space is spherical. In this chapter, we build upon these ideas to build a model that embeds ordinal data into the spherical latent space. Most our attention in this chapter focus on two key aspects. First, we carefully examine different types of ordinal structures that could be incorporated into our model. Second, we evaluate various configurations available in the link function of our spherical model using both simulated and real data sets.

The remainder of the chapter is organized as follows: Section 4.1 provides a brief overview of different ordinal models in the Euclidean space. Section 4.2 introduces our spherical latent factor model for ordinal data which employs the continuation-ratio formulation. In addition, we also specify four different configurations for the link function in our model. Section 4.3 discusses our computational approach

which is similar to those implemented in the previous two chapters. Section 4.4 surveys various robust metrics that work well in the case of unbalanced ordinal data which is common in survey applications. Section 4.5 compares the performance of our model with different configurations from which we select the best model to compare against the Euclidean model. Lastly, Section 4.6 concludes this chapter with a discussion.

4.1 Euclidean Ordinal Latent Factor Model

In this section, we review various standard latent factor models for ordinal data that implicitly rely on Euclidean embedding. In particular, we discuss models for ordinal data with or without *proportional odds* (PO or NPO) structure [88] in the context of latent variable modeling. In the literature, proportional odds structure is sometimes referred to as *parallel* structure [89].

Consider data consisting of independent multivariate ordinal observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$ associated with i *subjects*, where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,J})^T$ is a vector where each entry is associated with a different *item*, and $y_{i,j} \in \{1, 2, \ldots, L_j\}$. For example, in customer review applications, $y_{i,j}$ might represent the rating consumer (subject) i gives to (item) j in a scale of 1 star to 5 stars. On the other hand, in survey applications, $y_{i,j}$ might represent the psychometric scale of respondent (subject) i to (question) j using a typical five-level Likert scale such as strongly disagree, disagree, neutral, agree, strongly agree. The subscript j in L_j is necessary since the number of categories could vary across items. When $L_j = 2$, it corresponds to a binary response which has been studied extensively in the previous two chapters. In this chapter, we focus on the embedding of ordinal data and hence assume

$L_j \geq 3$.

4.1.1 Cumulative model

The PO and NPO cumulative models are formulated as follows,

$$\text{PO:} \qquad \Pr\left(Y_{i,j} \leq l \mid \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,\cdot}^T \boldsymbol{\beta}_i\right), \qquad l = 1, \ldots, L_j - 1, \tag{4.1}$$

$$\text{NPO:} \qquad \Pr\left(Y_{i,j} \leq l \mid \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \qquad l = 1, \ldots, L_j - 1, \tag{4.2}$$

where the cutoff points (intercepts) $\mu_{1,l}, \ldots, \mu_{J,l}$ as well as the bilinear terms $\boldsymbol{\alpha}_{j,\cdot}, \boldsymbol{\alpha}_{1,l}, \ldots, \boldsymbol{\alpha}_{J,l} \in \mathbb{R}^{K+1}$ and $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_I \in \mathbb{R}^{K+1}$ are all unknown and need to be estimated from the data, and G is a link function (typically the cumulative distribution function of logistic, normal or extreme value distribution which corresponds to logit, probit and complementary log-log model respectively). The class probability of category l is,

$$\begin{aligned}
&\Pr(Y_{i,j} = 1 \mid \boldsymbol{\theta}) = \Pr(Y_{i,j} \leq 1 \mid \boldsymbol{\theta}), \\
&\Pr(Y_{i,j} = l \mid \boldsymbol{\theta}) = \Pr(Y_{i,j} \leq l \mid \boldsymbol{\theta}) - \Pr(Y_{i,j} \leq l-1 \mid \boldsymbol{\theta}), \quad \text{for } l = 2, \ldots, L_j - 1, \\
&\Pr(Y_{i,j} = L_j \mid \boldsymbol{\theta}) = 1 - \Pr(Y_{i,j} \leq L_j - 1 \mid \boldsymbol{\theta}).
\end{aligned} \tag{4.3}$$

For a given (item) j, the proportional odds structure arises from associating $\boldsymbol{\beta}_i$ with the same slope $\boldsymbol{\alpha}_{j,\cdot}$ across all categories. The original terminology came from [90] where G is a CDF of standard logistic distribution (G^{-1} is logit link function).

More specifically,

$$\begin{aligned}&\text{Logit}\left(\Pr(Y_{i,j} \leq l \mid \boldsymbol{\beta}_1\right) - \text{Logit}\left(\Pr(Y_{i,j} \leq l \mid \boldsymbol{\beta}_2\right) \\ &= \log \frac{\Pr(Y_{i,j} \leq l \mid \boldsymbol{\beta}_1)/\Pr(Y_{i,j} > l \mid \boldsymbol{\beta}_1)}{\Pr(Y_{i,j} \leq l \mid \boldsymbol{\beta}_2)/\Pr(Y_{i,j} > l \mid \boldsymbol{\beta}_2)} = \boldsymbol{\alpha}_{j,\cdot}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2), \end{aligned} \tag{4.4}$$

and the odds of making response $Y_{i,j} \leq l$ at $\boldsymbol{\beta}_1$ are $\exp\left(\boldsymbol{\alpha}_{j,\cdot}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right)$ times the odds at $\boldsymbol{\beta}_2$. For a one-unit increase in the difference of $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$, the cumulative log odds ratio is increased by $\boldsymbol{\alpha}_{j,\cdot}$. In addition, the condition $\mu_{j,1} < \mu_{j,2} < \ldots \mu_{j,L_j-1}$ allows *stochastic ordering* of the category along an underlying latent linear continuum. The stochastic ordering property ensures that the class probabilities in Equation (4.3) are all non-negative. Furthermore, it allows $\boldsymbol{\alpha}_{j,\cdot}$ to be invariant to the choice of response categories (91, pg. 56) in the ordinal regression setting. For example, collapsing a five-level scale (strongly disagree, disagree, neutral,agree, strongly agree) to a three-level scale (disagree, neutral, agree) will leave $\boldsymbol{\alpha}_{j,\cdot}$ invariant. This unique feature of the PO cumulative model makes model comparison or meta-analysis from studies using different scales possible. However, in the scope of latent factor modeling setting when both $\boldsymbol{\alpha}_j$s and $\boldsymbol{\beta}_i$s are unknown, this invariance property only holds if the model is identifiable.

The cumulative model without proportional odds structure (NPO) is obtained by allowing the slope $\boldsymbol{\alpha}_{j,l}$ to be different for each category l in $1, \ldots, L_j - 1$. However, other than the usual parsimony argument, this model have a major limitation. The curves for different cumulative probabilities must eventually intersect due to non-parallel slope (90, pg. 155), which violates the stochastic ordering of the cumulative probabilities (91, pg. 76) and leads to negative class probability in equation (4.3). [92] proposes a Bayesian framework for logit cumulative model to address this problem by incorporating the stochastic ordering constraint into the

joint posterior and then employ a truncated sampling scheme to obtain posterior samples. In addition, they also present a model selection framework to choose between PO and NPO structure using reversible-jump MCMC. Alternatively, [93] suggests a hierarchical ordered probit model which uses positive incremental cutoff points through a nonlinear specification.

4.1.2 Adjacent-category model

The PO and NPO adjacent-category model are formulated as follows for $l = 1, \ldots, L_j - 1$,

$$\text{PO:} \quad \Pr\left(Y_{i,j} = l \mid Y_{i,j} = l \text{ or } Y_{i,j} = l+1, \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,\cdot}^T \boldsymbol{\beta}_i\right), \tag{4.5}$$

$$\text{NPO:} \quad \Pr\left(Y_{i,j} = l \mid Y_{i,j} = l, \text{ or } Y_{i,j} = l+1, \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \tag{4.6}$$

where the intercepts $\mu_{1,l}, \ldots, \mu_{J,l}$ as well as the bilinear terms $\boldsymbol{\alpha}_{j,\cdot}, \boldsymbol{\alpha}_{1,l}, \ldots, \boldsymbol{\alpha}_{J,l} \in \mathbb{R}^{K+1}$ and $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_I \in \mathbb{R}^{K+1}$ are again all unknown and need to be estimated from the data, and G is a link function. The class probability for category t is,

$$\pi_{i,j,t} = \Pr(Y_{i,j} = t) = \frac{\prod_{m=1}^{t-1}\left[1 - G(\eta_{i,j,m})\right]\prod_{n=t}^{L-1} G(\eta_{i,j,n})}{\sum_{t=1}^{L}\left[\prod_{m=1}^{t-1}\left[1 - G(\eta_{i,j,m})\right]\prod_{n=t}^{L-1} G(\eta_{i,j,n})\right]}, \text{ for } t = 1, \ldots, L_j, \tag{4.7}$$

where $\eta_{i,j,m} = \mu_{j,m} + \boldsymbol{\alpha}_{j,\cdot}^T \boldsymbol{\beta}_i$ in the PO model and $\eta_{i,j,m} = \mu_{j,m} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i$ in the NPO model.

Similar to the PO cumulative model, the proportional odds structure here can be

easily obtained if G is a CDF of standard logistic distribution,

$$\begin{aligned}
&\text{Logit}\left[\Pr\left(Y_{i,j}=l \mid Y_{i,j}=l \text{ or } Y_{i,j}=l+1, \boldsymbol{\beta}_1\right)\right] \\
&\quad -\text{Logit}\left[\Pr\left(Y_{i,j}=l \mid Y_{i,j}=l \text{ or } Y_{i,j}=l+1, \boldsymbol{\beta}_2\right)\right] \\
&\quad =\text{Logit}\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l}+\pi_{i,j,l+1}} \mid \beta_1\right)-\text{Logit}\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l}+\pi_{i,j,l+1}} \mid \beta_2\right) \\
&\quad =\log\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l+1}} \mid \beta_1\right)-\log\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l+1}} \mid \beta_2\right)=\boldsymbol{\alpha}_{j,\cdot}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2), \qquad (4.8)
\end{aligned}$$

Similarly, for a one-unit increase in the difference of $\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$, the adjacent-category log odds ratio is increased by $\boldsymbol{\alpha}_{j,\cdot}$.

Unlike the NPO cumulative model, the NPO adjacent-category model is guaranteed to have non-negative probability for all categories by construction. In addition, the NPO adjacent category model is *permutation invariant* (90, p116) and has a well-known connection with the baseline category model under the logit link function (91, p91) for nominal data. In other words, adjacent category NPO model is not a ordinal model and is best suited to situations in which there is ambiguity in the category order. These features are often associated with the traditional regression setting, but also are valid in ordinal latent factor modeling setting. Here we show how the adjacent-category NPO model is connected to the baseline logit model (88, p293). For $u<v$ and $u,v\in(1,\ldots L_j)$,

$$\begin{aligned}
\log\left(\frac{\pi_{i,j,u}}{\pi_{i,j,v}}\right)&=\log\left(\frac{\pi_{i,j,u}}{\pi_{i,j,u+1}}\right)+\log\left(\frac{\pi_{i,j,u+1}}{\pi_{i,j,u+2}}\right)+\cdots+\log\left(\frac{\pi_{i,j,v-1}}{\pi_{i,j,v}}\right) \\
&=\left\{\sum_{l=u}^{v-1}\mu_{j,l}\right\}+\left\{\sum_{l=u}^{v-1}\boldsymbol{\alpha}_{j,l}^T\right\}\boldsymbol{\beta}_i=\mu_{j,u}^*+(\boldsymbol{\alpha}_{j,u}^*)^T\boldsymbol{\beta}_i. \qquad (4.9)
\end{aligned}$$

Evidently, this formulation is equivalent to a baseline logit model if v represents the baseline category. Note that in the case of adjacent-category PO model,

$\sum_{l=u}^{v-1} \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i$ in (4.9) becomes $\boldsymbol{\alpha}_{j,\cdot}^T (v-u)$, which clearly shows that the PO model recognizes the ordering of the response since the effect is proportional to the the distance between categories.

4.1.3 Continuation-ratio model

The PO and NPO continuation-ratio model are formulated as follows,

$$\text{PO:} \quad \Pr\left(Y_{i,j} = l \mid Y_{i,j} \geq l, \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,\cdot}^T \boldsymbol{\beta}_i\right), \quad l = 1, \ldots, L_j - 1, \tag{4.10}$$

$$\text{NPO:} \quad \Pr\left(Y_{i,j} = l \mid Y_{i,j} \geq l, \boldsymbol{\theta}\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \quad l = 1, \ldots, L_j - 1, \tag{4.11}$$

where the intercepts $\mu_{1,l}, \ldots, \mu_{J,l}$ as well as the bilinear terms $\boldsymbol{\alpha}_{j,\cdot}, \boldsymbol{\alpha}_{1,l}, \ldots, \boldsymbol{\alpha}_{J,l} \in \mathbb{R}^d$ and $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_I \in \mathbb{R}^K$ are again all unknown and need to be estimated from the data, G is again a link function similar. The class probability of category is,

$$\begin{aligned}
&\Pr(Y_{i,j} = 1 \mid \boldsymbol{\theta}) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,*}^T \boldsymbol{\beta}_i\right), \\
&\Pr(Y_{i,j} = l \mid \boldsymbol{\theta}) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,*}^T \boldsymbol{\beta}_i\right) \prod_{t=1}^{l-1} \left[1 - G\left(\mu_{j,t} + \boldsymbol{\alpha}_{j,*}^T \boldsymbol{\beta}_i\right)\right], \ l = 2, \ldots, L_j - 1, \\
&\Pr(Y_{i,j} = L_j \mid \boldsymbol{\theta}) = \prod_{t=1}^{L_j - 1} \left[1 - G\left(\mu_{j,t} + \boldsymbol{\alpha}_{j,*}^T \boldsymbol{\beta}_i\right)\right],
\end{aligned} \tag{4.12}$$

where $\boldsymbol{\alpha}_{j,*}^T = \boldsymbol{\alpha}_{j,\cdot}^T$ and $\boldsymbol{\alpha}_{j,*}^T = \boldsymbol{\alpha}_{j,t}^T$ corresponds to the PO and NPO continuation-ratio models respectively.

Similar to the PO cumulative model and the PO adjacent-category model, the proportional odds structure here once again can be easily observed if G is a CDF

of standard logistic distribution,

$$\begin{aligned}
&\text{Logit}\left[\Pr\left(Y_{i,j}=l \mid Y_{i,j}\geq l,\boldsymbol{\beta}_1\right)\right]-\text{Logit}\left[\Pr\left(Y_{i,j}=l \mid Y_{i,j}\geq l,\boldsymbol{\beta}_2\right)\right]\\
&\quad=\log\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l+1}+\cdots+\pi_{i,j,L_j}} \mid \beta_1\right)-\log\left(\frac{\pi_{i,j,l}}{\pi_{i,j,l+1}+\cdots+\pi_{i,j,L_j}} \mid \beta_2\right)\\
&\quad=\boldsymbol{\alpha}_{j,\cdot}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2), \qquad (4.13)
\end{aligned}$$

Similarly, for a one-unit increase in the difference of $\boldsymbol{\beta}_1-\boldsymbol{\beta}_2$, the continuation-ratio log odds ratio is increased by $\boldsymbol{\alpha}_{j,\cdot}$.

Unlike the NPO cumulative models, the NPO continuation-ratio models are guaranteed to have non-negative probability for all categories by construction.

4.1.4 Alternative construction of PO and NPO models through random utility functions

The class of factor analysis models for ordinal data can also be constructed through the use of random utility function [12, 13] similar to the factor models discussed in previous chapters. To develop such construction, we need to assume that each subject i has associated with it a position $\boldsymbol{\beta}_i \in \mathbb{R}^{K+1}$ (which can be interpreted as representing their preferences over a set of unobserved item characteristics) and associate with each item L_j-1 pairs of positions, $\boldsymbol{\psi}_{j,l} \in \mathbb{R}^{K+1}$ (corresponding to a positive response for category l, i.e., $Y_{i,j}\geq l$ in the cumulative model) and $\boldsymbol{\zeta}_{j,l} \in \mathbb{R}^{K+1}$ (corresponding to a negative one for category l, i.e., $Y_{i,j}<l$ in the cumulative model). Given these positions, individuals make their choice about category l of item j based on the relative value of two random quadratic utilities,

for $l = 1, \ldots, L_j - 1$,

$$U_+^l(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i) = -\left\|\boldsymbol{\psi}_{j,l} - \boldsymbol{\beta}_i\right\|^2 + \epsilon_{i,j}^l, \quad U_-^l(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i) = -\left\|\boldsymbol{\zeta}_{j,l} - \boldsymbol{\beta}_i\right\|^2 + \nu_{i,j}^l, \quad (4.14)$$

where $\epsilon_{i,j}^l$ and $\nu_{i,j}^l$ represent random shocks to the utilities and $\upsilon_{i,j}^l = \nu_{i,j}^l - \epsilon_{i,j}^l$ are independently distributed for all i, j and l and have cumulative distribution function $G_{j,l}(x) = G(x/\sigma_{j,l})$. Under these assumptions, we can conveniently construct the NPO cumulative model , NPO adjacent-category model and the NPO continuation-ratio model through such random utility framework, for $l = 1, \ldots, L_j - 1$,

Cumulative Model :

$$\Pr\left(Y_{i,j} \geq l \mid \boldsymbol{\theta}\right) = \Pr\left(U_+^l\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i\right) > U_-^l\left(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \quad (4.15)$$

Adjacent-category Model :

$$\begin{aligned}\Pr\left(Y_{i,j} = l \mid Y_{i,j} = l \text{ or } Y_{i,j} = l+1, \boldsymbol{\theta}\right) &= \Pr\left(U_+^l\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i\right) > U_-^l\left(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right) \\ &= G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \quad (4.16)\end{aligned}$$

Continuation-ratio Model :

$$\Pr\left(Y_{i,j} = l \mid Y_{i,j} \geq l, \boldsymbol{\theta}\right) = \Pr\left(U_+^l\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i\right) > U_-^l\left(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right) = G\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,l}^T \boldsymbol{\beta}_i\right), \quad (4.17)$$

where $\boldsymbol{\alpha}_{j,l} = 2(\boldsymbol{\psi}_{j,l} - \boldsymbol{\zeta}_{j,l})/\sigma_{j,l}$ and $\mu_{j,l} = -(\boldsymbol{\psi}_{j,l} - \boldsymbol{\zeta}_{j,l})^T(\boldsymbol{\psi}_{j,l} + \boldsymbol{\zeta}_{j,l})/\sigma_{j,l}$. In addition, we can obtain the PO models if $(\boldsymbol{\psi}_{j,l} - \boldsymbol{\zeta}_{j,l})$s are fixed across categories while $(\boldsymbol{\psi}_{j,l} + \boldsymbol{\zeta}_{j,l})$s are allowed to vary. The identifiability issues associated with these Euclidean models are similar to those discussed in Section 1.2.

4.1.5 Bayesian Euclidean ordinal latent factor model

For reasons discussed next in Section 4.2, we employ the continuation-ratio construction of the Euclidean ordinal latent factor model under the Bayesian framework. In particular, we assume G in (4.17) to be the CDF of the normal distribution and thus we obtain a probit continuation-ratio model. This formulation leads to class probability,

$$\begin{aligned}
\pi_{i,j}^{1} &= \Pr(Y_{i,j} = 1 \mid \boldsymbol{\theta}) = \Phi\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,*}^{T}\boldsymbol{\beta}_i\right), \\
\pi_{i,j}^{l} &= \Pr(Y_{i,j} = l \mid \boldsymbol{\theta}) = \Phi\left(\mu_{j,l} + \boldsymbol{\alpha}_{j,*}^{T}\boldsymbol{\beta}_i\right) \prod_{t=1}^{l-1}\left(1 - \Phi\left(\mu_{j,t} + \boldsymbol{\alpha}_{j,*}^{T}\boldsymbol{\beta}_i\right)\right), \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad l = 2, \ldots, L_j - 1, \\
\pi_{i,j}^{L_j} &= \Pr(Y_{i,j} = L_j \mid \boldsymbol{\theta}) = \prod_{t=1}^{L_j-1}\left(1 - \Phi\left(\mu_{j,t} + \boldsymbol{\alpha}_{j,*}^{T}\boldsymbol{\beta}_i\right)\right), \qquad (4.18)
\end{aligned}$$

where $\boldsymbol{\alpha}_{j,*}^{T} = \boldsymbol{\alpha}_{j,\cdot}^{T}$ and $\boldsymbol{\alpha}_{j,*}^{T} = \boldsymbol{\alpha}_{j,t}^{T}$ corresponds to PO and NPO continuation-ratio model respectively, and leads to likelihood of the form,

$$\Pr(\boldsymbol{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\beta}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\pi_{i,j}^{l})^{z_{i,j}^{l}}, \qquad (4.19)$$

where $z_{i,j}^{l}$ is an indicator variable and is 1 if and only if $Y_{i,j} = l$. We assume $\mu_{j,l} \sim N(0,1)$, $\alpha_{j,*} \sim N(0,1)$, and $\beta_{j,k} \sim N(0, 1/k^2)$. Posterior inference for both PO and NPO Euclidean continuation-ratio model is carried out using HMC (see Section 1.3.1).

4.2 Bayesian Spherical Ordinal Latent Factor Model

In this section we focus on the NPO continuation-ratio model, which is both general and flexible, avoids the negative probability issues of the NPO cumulative, and the prohibiting complexity of [92] and [93].

Following the same formulation from Section 3.1, we can modify the utility functions in (14) by embedding the points $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\zeta}$ in a spherical manifold by constructing the utility function through the geodesic distance, leading to

$$U_+(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i) = -\left\{\rho\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i\right)\right\}^2 + \epsilon_{i,j}^l, \quad U_-(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i) = -\left\{\rho\left(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right\}^2 + \nu_{i,j}^l, \tag{4.20}$$

where the errors $\epsilon_{i,j}^l$ and $\nu_{i,j}^l$ are such that their differences $\upsilon_{i,j}^l = \nu_{i,j}^l - \epsilon_{i,j}^l$ are independent for all i, j and l, and have cumulative distribution function $G_{j,l}$. The NPO spherical continuation-ratio model constructed under the random utility framework leads to the following class probability,

$$\begin{aligned}
\theta_{i,j}^1 &= \Pr(Y_{i,j} = 1 \mid \boldsymbol{\theta}) = G_{j,1}\left(e\left(\boldsymbol{\psi}_{j,1}, \boldsymbol{\zeta}_{j,1}, \boldsymbol{\beta}_i\right)\right), \\
\theta_{i,j}^l &= \Pr(Y_{i,j} = l \mid \boldsymbol{\theta}) = G_{j,l}\left(e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right) \prod_{t=1}^{l-1}\left[1 - G_{j,l}\left(e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right)\right], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } l = 2, \ldots, L_j - 1, \\
\theta_{i,j}^{L_j} &= \Pr(Y_{i,j} = L_j \mid \boldsymbol{\theta}) = \prod_{t=1}^{L_j-1}\left[1 - G_{j,L_j}\left(e\left(\boldsymbol{\psi}_{j,L_j}, \boldsymbol{\zeta}_{j,L_j}, \boldsymbol{\beta}_i\right)\right)\right],
\end{aligned} \tag{4.21}$$

where $e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right) = \left\{\rho\left(\boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right\}^2 - \left\{\rho\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\beta}_i\right)\right\}^2$ for $l = 1, \ldots, L_j$, and leads

to likelihood of the form,

$$\Pr(\boldsymbol{Y} \mid \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\theta_{i,j}^{l}\right)^{z_{i,j}^{l}}, \tag{4.22}$$

where $z_{i,j}^{l}$ is an indicator variable and is 1 only if $Y_{i,j} = l$. For reasons discussed in Section 3.1.1, we adopt the same hyperspherical coordinates which simplifies the computation of the geodesic distance $\rho_K(\boldsymbol{x}, \boldsymbol{z}) = \arccos\left(\boldsymbol{x}^T \boldsymbol{z}\right)$ and facilitates the development of our computational approaches.

We use the spherical von Mises distribution (see Equation (3.5)) proposed in the previous chapter as priors for the $\boldsymbol{\psi}_{j,l}$s, $\boldsymbol{\zeta}_{j,l}$s and $\boldsymbol{\beta}_i$s. In particular, we set

$$\boldsymbol{\psi}_i \sim \text{SvM}(\tau, 2^2\tau, 3^2\tau, \ldots, K^2\tau) \tag{4.23}$$

$$\boldsymbol{\zeta}_{j,l} \sim \text{SvM}(\tau, 2^2\tau, 3^2\tau, \ldots, K^2\tau) \tag{4.24}$$

$$\boldsymbol{\beta}_{j,l} \sim \text{SvM}(\omega, 2^2\omega, 3^2\omega, \ldots, K^2\omega) \tag{4.25}$$

independently across all $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $l = 1, \ldots, L_j$. The corresponding density of the Hausdorff measure with respect to the uniform distribution on the sphere associated the spherical von Mises distributions is given in Equation (3.7).

The likelihood function for the spherical factor model discussed in Section 4.2 is invariant to simultaneous rotations of all latent positions. The identifiability issues associated $\boldsymbol{\psi}_{j,l}$s, $\boldsymbol{\zeta}_{j,l}$s, and $\boldsymbol{\beta}_i$s in the posterior distribution are addressed similarly as in Section 3.1.4.

We discussed the connection of our proposed spherical model to the Euclidean model for binary data analysis in Section 3.1.3. Such connection still holds for the

ordinal models discussed in this chapter. In particular, the K-dimensional NPO Euclidean continuation-ratio model with probit link discussed in Section 4.1.5 can be seen as a limiting case of our NPO spherical continuation-ratio model on $\mathcal{S}^K$ with symmetric link function (configuration 2 and 4 in Table 4.1).

Next we focus on our discussion on the link function specifications and some hyperpriors which are different than those discussed in the previous chapter.

4.2.1 Link function specifications

The link function $G_{j,l}$ must account for the fact that the function $e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)$ has as its range the interval $[-\pi^2, \pi^2]$. We again choose the cumulative distribution of a shifted and scaled beta distribution,

$$G_{j,l}(z) = \int_{-\pi^2}^{z} \frac{1}{2\pi^2} \frac{\Gamma(a_{j,l} + b_{j,l})}{\Gamma(a_{j,l})\Gamma(b_{j,l})} \left(\frac{\pi^2 + z}{2\pi^2}\right)^{a_{j,l}-1} \left(\frac{\pi^2 - z}{2\pi^2}\right)^{b_{j,l}-1} dz, \quad z \in [-\pi^2, \pi^2]. \tag{4.26}$$

However, there are several configurations of the parameters for this link function. If we let $a_{j,l} = b_{j,l} = \kappa_{j,l}$ as in previous chapters, $G_{j,l}$ becomes symmetric. Because, $\boldsymbol{\psi}_{j,l}$ and $\boldsymbol{\zeta}_{j,l}$ are assigned the same prior distributions, it is clear from a simple symmetry argument that, under this choice of $a_{j,l}$ and $b_{j,l}$ for a symmetric $G_{j,l}$, $\mathrm{E}(G_{j,l}\left(e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right)) = 1/2$ for all values of ω, τ and $\kappa_{j,l}$. Therefore, $\mathrm{E}(\theta_{i,j}^l) = \left(\frac{1}{2}\right)^l$ and the prior class probability distribution of $\theta_{i,j}^l$ is highly skewed to the right, leading to increasingly smaller probabilities for the larger categories. Therefore, we also consider the asymmetric parameterization to allow a more flexible prior class probability distribution.

If we let $a_{j,1} = a_{j,2} = \cdots = a_{j,L_j} = s_j\kappa_j$ and $b_{j,1} = b_{j,2} = \cdots = b_{j,L_j} = \kappa_j$ ($s_j = 1$ corresponds to a symmetric link and $s_j \neq 1$ for a asymmetric link), the dispersion of $G_{j,l}$ for each category l is constrained to be identical. We denote such parameterization as “Tied" structure throughout the rest of this chapter. Using a “Tied" structure ensures the same dispersion as a priori across all categories of (item) j. Models without such structure are more flexible in the presence of heterogeneous error distributions across ordinal categories. Therefore, we can specify four various parameter configurations for $G_{j,l}$, which are summarized in the following table ordered by the number of parameters in a non-decreasing order (configuration 3 can have the same number of parameters as configuration 2 if and only if $L_j = 3$ for all j),

Table 4.1: Configurations of $G_{j,l}$

Configuration	Asymmetric	Tied	Number of Parameters	$a_{j,l}$	$b_{j,l}$
1	False	True	J	κ_j	κ_j
2	True	True	$2J$	$s_j\kappa_j$	κ_j
3	False	False	$\sum_{j=1}^{J}(L_j - 1)$	$\kappa_{j,l}$	$\kappa_{j,l}$
4	True	False	$2\sum_{j=1}^{J}(L_j - 1)$	$s_{j,l}\kappa_{j,l}$	$\kappa_{j,l}$

Note that configuration 2 generalizes configuration 1, configuration 3 generalizes configuration 1 while configuration 4 generalizes all other configurations. We will compare the spherical models under these configurations in Section 4.5 using both simulated and real data sets.

4.2.2 Hyperpriors

Completing the specification of the model requires that we assign hyperpriors to ω, τ and the parameters of link function in Table 4.1. The precisions ω and τ are assigned independent Gamma distributions, $\omega \sim \text{Gam}(a_\omega, b_\omega)$ and $\tau \sim \text{Gam}(a_\tau, b_\tau)$.

The concentration parameters associated with the symmetric link function are assumed to be conditionally independent and given a common prior, $\kappa_j \sim \text{Gam}(c, \lambda)$ and $\kappa_{j,l} \sim \text{Gam}(c, \lambda)$ respectively, where λ is in turn given a conditionally conjugate Gamma hyperprior, $\lambda \sim \text{Gam}(a_\lambda, b_\lambda)$. The parameters a_ω, b_ω, a_τ, b_τ, a_λ, b_λ and c for these hyperpriors are assigned to strongly favor configurations in which $\beta_{i,1} \in [-\pi/2, \pi/2]$ (which is consistent with the assumption that a Euclidean continuation-ratio model is approximately correct). We set these hyperparameters the same way as discussed in Section 3.1.5.

Similarly, On the other hand, the parameters s_j and $s_{j,l}$, which controls asymmetricity of the link function, are assumed to be conditionally independent and are also given a common prior $\text{Gam}(a_s, b_s)$. Recall that $\text{E}(G_{j,l}\left(e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right)) = 1/2$ if symmetric link function is used, which induces increasingly smaller probabilities for the later categories. Hence, hyperparameters a_s and b_s are specified such that it has high prior probability around 1 for which the asymmetric model becomes the symmetric model, and high prior probability favoring $\text{E}(G_{j,l}\left(e\left(\boldsymbol{\psi}_{j,l}, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\beta}_i\right)\right)) < 1/2$. For these reasons, we suggest using $a_s = 4$ and $b_s = 2$, which is used throughout the rest of this chapter.

4.3 Computation

The posterior distribution for the spherical factor model is again analytically intractable. Similar to the previous two chapters, inference for the model parameters is carried out using a hybrid that combines Gibbs sampling, random walk Metropolis-Hastings and Hamiltonian Monte Carlo steps to generate samples from the full conditional distributions of each parameter. The simplest steps correspond to sampling the parameters ω, τ, λ, and parameters associated with link function. In particular, we sample λ from its Gamma full conditional posterior distribution, and sample ω, τ and parameters associated with the link function in Table 4.1 using random walk Metropolis Hastings with log-Gaussian proposals. The variance of the proposals for these steps are tuned so that the acceptance rate is roughly 40%. On the other hand, for sampling the latent positions we again employ the Geodesic Hamiltonian Monte Carlo (GHMC) algorithm described in Section 1.3.3.

As an example, consider the step associated with updating $\boldsymbol{\beta}_i$, the latent factor for individual i. Denoting the associated coordinates in $\mathbb{R}^{K+1}$ by $\boldsymbol{x}_{\beta_i}$s (recall Equation 3.3), the density of the Hausdorff measure associated with the full conditional distribution is given by

$$p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \cdots) \propto \left[\prod_{j=1}^{J}\prod_{l=1}^{L_j}(\theta_{i,j}^{l}\left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}\right))^{z_{i,j}^{l}},\right]\left[\exp\left\{\omega\frac{x_{\beta_i,1}}{\sqrt{x_{\beta_i,1}^2+x_{\beta_i,2}^2}}\right\}\right]$$
$$\exp\left\{-\sum_{k=2}^{K}k^2\omega\left(2\frac{x_{\beta_i,k+1}^2}{\sum_{t=1}^{k+1}x_{\beta_i,t}^2}-1\right)\right\}\left[\frac{1}{\prod_{k=1}^{K}\left(\sum_{t=1}^{k+1}x_{\beta_i,t}^2\right)^{\frac{1}{2}}}\right], \qquad \boldsymbol{x}_{\beta_i}^T\boldsymbol{x}_{\beta_i}=1,$$

where $\boldsymbol{x}_{\beta_i} = (x_{\beta_i,1}, x_{\beta_i,2}, \ldots, x_{\beta_i,K+1})$ and $\theta_{i,j}^{l}$ is defined in (4.21). Then, given tuning parameters L and ϵ, the GHMC preceeds the same way as that described in Section 3.2.

Recall that in Appendix A.4, we derived the gradient of the logarithm of the Hausdorff measure pertain to the same prior used in this chapter. Detailed expressions for the Hausdorff measures associated with the full conditional distributions of the $\boldsymbol{x}_{\beta_i}$s, $\boldsymbol{x}_{\psi_j}$s and $\boldsymbol{x}_{\zeta_j}$s, as well as their corresponding gradients, can be found in Appendix A.5. Similar to the previous two chapters, we periodically "jitter" the step sizes and the number of leap steps (e.g., see 77, pg. 306) in our experiments. This approach greatly improved the mixing of the algorithm. The specific range in which ϵ and L move for each (group of) parameter and each data set is again selected to target an average acceptance probability between 60% and 90% [65, 67].

4.4 Robust Metrics for Ordinal Data

In binary data analysis, Information Criterion such as WAIC, DIC, AIC/BIC, or predicted test accuracy based measures could be employed to perform model selections. However, such metrics may not work well in the presence of unbalanced data which is prevalent in ordinal data analysis. For example, a trivial model that assigns all responses to a single dominant class might outperform carefully constructed models. Therefore, we also employ robust measures of predicted accuracy including *AMAE* [94] and *MMAE* [95] to evaluate model performance. In addition, we also propose an alternative robust metric, *AACC*. These metrics, along with traditional metrics including *ACC* and *ACC1*, are defined as follows:

1. *ACC* denotes the percentage of correctly predicted labels among all labels.

$$ACC = \frac{\sum_{n=1}^{N} z_n}{N}, \tag{4.27}$$

where z_n denotes an indicator variable and it is 1 if the model correctly predicts the label of y_n.

2. *ACC1* represents the accuracy within 1 category that relaxes the above measure by allowing category $l-1$ (true label is l) to be also considered as "correct". For example, if the true label is 3 stars, then predicted label of either 3 or 2 stars would both be considered as "correct",

$$ACC1 = \frac{\sum_{n=1}^{N} z_n^*}{N} \tag{4.28}$$

where z_n^* denotes an indicator variable and it is 1 if model correctly predicts the label of y_n within 1 category.

3. *AMAE* is the average mean absolute error across categories. It is robust in the presence of unbalanced data, and becomes the traditional mean absolute error (*MAE*) if the data is balanced.

$$AMAE = \frac{1}{L}\sum_{l=1}^{L}\left(\frac{\sum_{u=1}^{N_l}|\hat{y_u^l} - y_u^l|}{N_l}\right), \tag{4.29}$$

4. *MMAE*, an alternative to *AMAE*, is the maximum mean absolute error across categories,

$$MMAE = \max_{l}\left(\frac{\sum_{u=1}^{N_l}|\hat{y_u^l} - y_u^l|}{N_l}\right), \tag{4.30}$$

where y_u^l denotes the true response corresponds to category l and N_l represents the associated total number of responses in that category while $\hat{y_u^l}$ is the predicted label.

5. *AACC*, which is similar to the idea of *AMAE* and *MMAE*, is the average

mean accuracy across categories proposed in this chapter. It is a robust measure that generalizes ACC if the data is balanced,

$$AACC : \frac{1}{L}\sum_{l}^{L}\left(\frac{\sum_{u=1}^{N_l} z_u^l}{N_l}\right). \tag{4.31}$$

where z_u^l denotes an indicator variable for response y_u^l belonging to category l and it is 1 if model correctly predicts the label of y_u^l.

To summarize, higher ACC, $ACC1$, $AACC$ and lower $AMAE$, $MMAE$ suggests a better overall fit. In the sequel, we refer to ACC and $ACC1$ as the traditional measures and the other as robust measures.

4.5 Illustrations

In this section, we illustrate the performance of the proposed models using both simulated and real data sets. We compare all four versions of the spherical model described in Table 4.1, as well as the PO and NPO Euclidean models from Section 4.1.5. Our objective is threefold: First, we evaluate and compare spherical models under four different link functions specified in Table 4.1. Secondly, we compare the spherical model with its Euclidean counterpart. Finally, we aim to evaluate our ability to select optimal model dimension using the Deviance Information Criteria (DIC) as well as the various performance metrics discussed in Section 4.4. To evaluate these performance metrics, we randomly select 5% of the data as the test set.

Computation for both the spherical and Euclidean models was carried out using HMC algorithms. In all of the analyses involving spherical models, the number

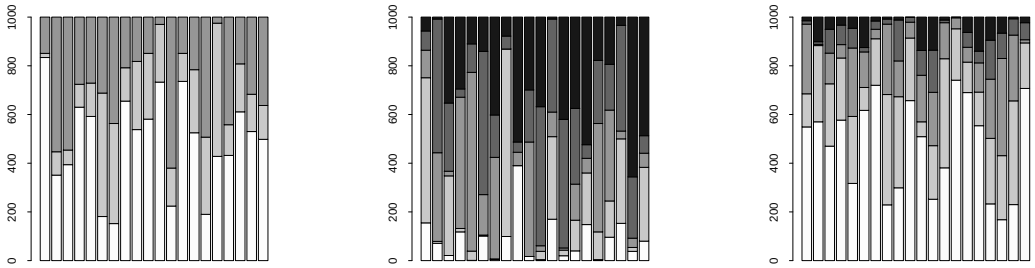
of leaps used in the HMC steps is randomly selected from a discrete uniform distribution between 1 and 10 or 5 and 10 every 50 samples. Similarly, the leap sizes are drawn from uniform distribution on $(0.01, \epsilon_\beta)$ for each $\boldsymbol{\beta}_i$ where $\epsilon_\beta \in [0.05, 0.2]$, and from a uniform distribution on $(0.005, \epsilon_\zeta)$ or $(0.01, \epsilon_\psi)$, for each $\boldsymbol{\zeta}_{j,l}$ and $\boldsymbol{\psi}_{j,l}$ where $\epsilon_\beta, \epsilon_\psi \in [0.01, 0.1]$. All inferences presented in this section are based on 20,000 samples obtained after convergence of the Markov chain Monte Carlo algorithm. For the simulation study, the length of burn in period are around 15,000 iterations. On the other hand, the length of the burn in period is 20,000 and 35,000 iterations for ASES and BEPS data respectively. We also used HMC to generate posterior samples from the PO and NPO models. The number of leaps used in the HMC steps is randomly selected from a discrete uniform distribution between 1 and 10 Convergence was checked by monitoring the value of the log-likelihood function, both through visual inspection of the trace plot, and by comparing multiple chains using the procedure in [78].

4.5.1 Simulation Study

We conducted a simulation study including three distinct scenarios to evaluate our spherical model under various link functions. Each simulated data set consists of $I = 1000$ subjects and $J = 20$ items. In the first two scenarios, the data is simulated from a spherical factor model under configuration 1 and 4 in Table 4.1 on $\mathcal{S}^3$ and $\mathcal{S}^2$ respectively. Recall that configuration 1 represents the most parsimonious version of the link function while configuration 4 has the most parameters. In both of these cases, the item-specific latent positions, as well as the subject-specific latent positions, are sampled from spherical von-Mises distributions where all component-wise precisions are equal to 2. The third data set

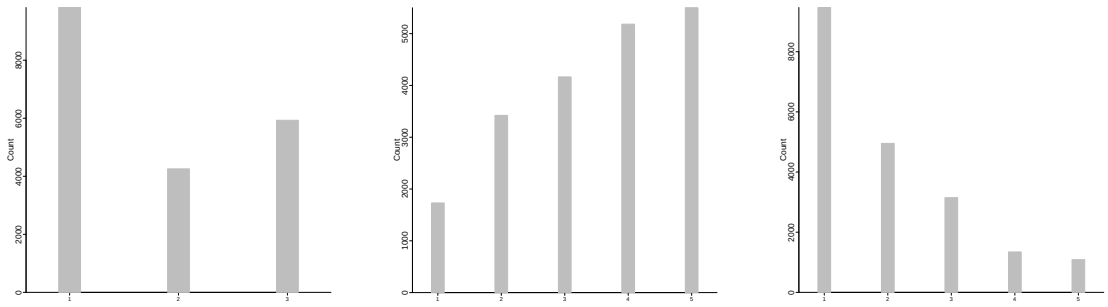
is simulated from a NPO Euclidean model on $\mathbb{R}^3$ in which the latent positions are generated from standard Gaussian distributions. In terms of the features of the items, the first simulated data consists of three-category items only while the later two simulated data sets consist solely of five-category items. In the second scenario, the scale parameters $s_{j,1}, s_{j,2}, s_{j,3}$ and $s_{j,4}$ which brings the asymmetricity to the model are sampled from a uniform distribution on $(1.7, 1.9)$, $(1.4, 1.6)$, $(1.1, 1.3)$, and $(0.9, 1.1)$ respectively to favor the case in which later categories have more responses.

We present the marginal distribution of the answers to each item as well as the discrete distribution of the data for these simulated data sets in Figure 4.1 and 4.2. All three data sets display some degree of imbalance by construction. In particular, simulated data 1 and 3 show a higher concentration in the early categories, while simulated data 2 favors the later categories.



(a) Simulation data 1 **(b)** Simulation data 2 **(c)** Simulation data 3

Figure 4.1: Marginal distribution of the answers to each of the 20 items in each simulated dataset. The ordinal scale is represented by the shades of gray from the lightest to the darkest.



(a) Simulated data 1 **(b)** Simulated data 2 **(c)** Simulated data 3

Figure 4.2: Overall frequency of responses for each simulated dataset.

We start by fitting spherical and Euclidean models of varying dimensions in all scenarios. The values of the DIC for each model and dimension are shown in Figure 4.4. In the first two scenarios, the best model selected by DIC coincides with the true data generating model. Moreover, the optimal NPO model in each of these two scenarios has a dimension that is one unit higher than the true spherical dimension, a result that is consistent with those in Section 3.3.1. On the other hand, in the third scenario, all spherical models achieve similar or slightly better performance than the true data generating (Euclidean) model. More specifically, in scenario 1, the optimal model identified by DIC correspond to a two-dimensional spherical model with a tied and symmetric link function. Nonetheless, we can see from the first column of Figure 4.3 that the first two dimensions of the $\boldsymbol{\beta}_i$s recovered by the spherical models (which, in every case, capture over 95% of the variability of the latent traits) are similar across all spherical models. Hence, while the more complex model is (reasonably) not preferred by DIC, the results suggests that inferences for the underlying latent traits will not be dramatically affected by the use of a link function that is more complex that you would really need. On the other hand, in scenario 2, DIC correctly identifies a two-dimensional spherical model with untied and asymmetric link functions as the best model for

the model. Furthermore, the second column of Figure 4.3 shows that $\boldsymbol{\beta}_i$s estimated from the correct model have a much higher variance than those estimated from the simpler models. This suggests that specifying a link function that is not flexible enough has serious implications, not only in terms of fit-complexity tradeoffs, but also in terms of the ability of the model to accurately recover the value of the latent traits. Finally, in scenario 3, DIC selects the optimal spherical models and data generating NPO Euclidean model all at the right dimension and the optimal spherical models achieve equivalent or better performance as the data generating Euclidean model. While slightly surprising, this result might be explained by the fact that the since the NPO Euclidean model is a limiting case for these spherical models. To conclude, note that the PO Euclidean model is under-parameterized and hence does not perform well in all of these scenarios.

Computing the DIC for each model and dimension as we just did is computationally intensive. An alternative approach to dimension selection is based on fitting a single high-dimensional model, and investigating the behavior of the principal nested sphere (PNS) decomposition of the estimated spherical latent space (recall Section 3.3.1). The principal nested spheres decomposition is implemented with a fixed radius of 1 in this chapter. Figure 4.5 employs such an approach for our three simulation scenarios and confirms several key observations from the DIC results discussed earlier. In particular, in scenario 1, all spherical models achieve an elbow at the correct dimension with almost identical variance decomposition. In scenario 2, only the data generating model retains an elbow at the true dimension while the remaining models require an additional dimension. In addition, the elbow from the data generating model is much sharper than its counterparts. In scenario 3, the variance of each spherical model plateaus at the same (correct) dimension.

Another approach to dimension selection as well as model comparison is also based on fitting a single high-dimensional model. In particular, we compare the nested and full model performance using metrics outlined in Section 4.4. Recall from Section 3.1.1, the nested model can be easily obtained from the full model through an recursive procedure by zeroing out the higher dimensions of the latent positions. The nested and the full model results are shown in Figure 4.6, Figure 4.7 and Figure 4.8 respectively. In the first scenario, the results demonstrate that all spherical models deliver comparable performance and each retains a sharp elbow at the underlying truth in both traditional and robust measures. In the second scenario, only the true data generating spherical model obtains a clear elbow at the underlying truth while the rest of the spherical models require an additional dimension to achieve their best performance respectively. In the third scenario, spherical models and the NPO Euclidean model all retain a clear elbow at the right dimension. Furthermore, the spherical models perform significantly better in terms of the robust measures than the data generating NPO Euclidean model. On the other hand, the PO Euclidean model once again does not show good performance in any of these scenarios. All of these observations resonates with the DIC and PNS results.

To summarize, the simulation study conducted suggests that we could use the nested model results computed from the full model to select the optimal latent dimension as well as perform model comparison without fitting models of varying dimensions.

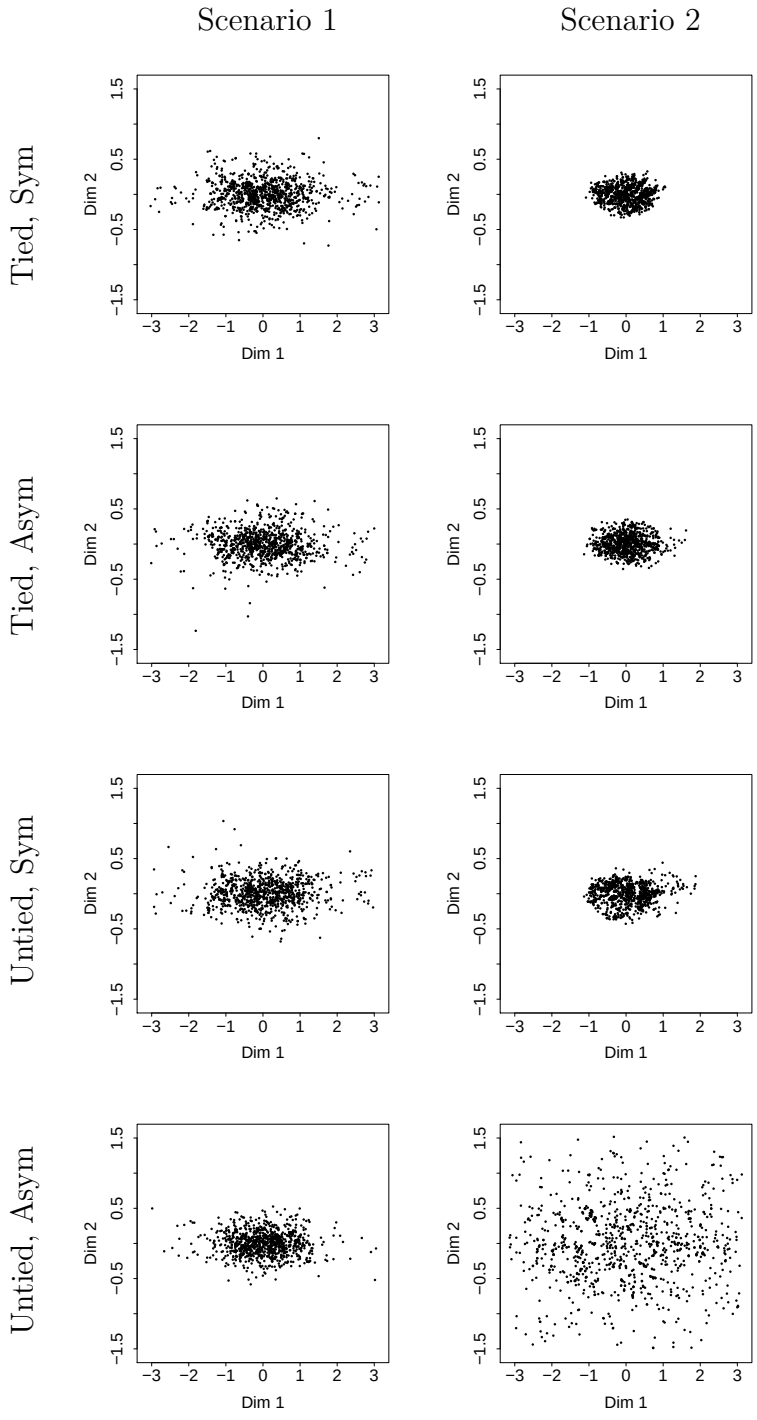
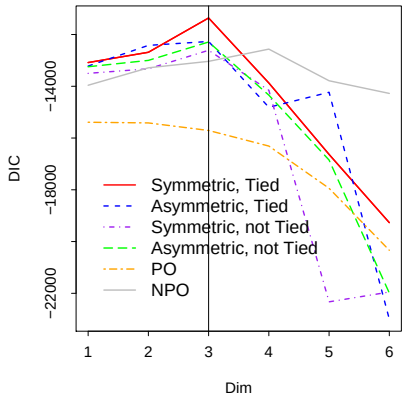
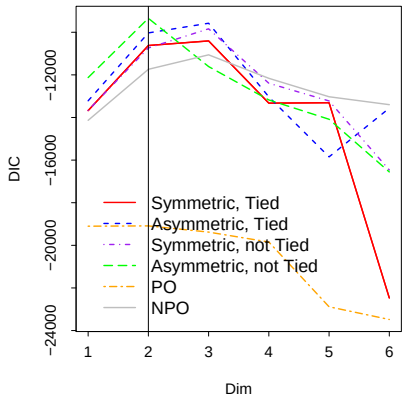


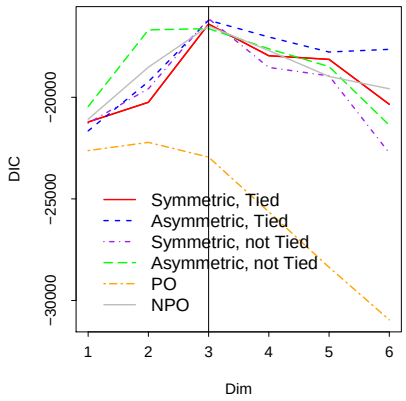
Figure 4.3: Pairwise plots of the first two dimensions for $\boldsymbol{\beta}_i$ in the simulation study.



(a) Scenario 1

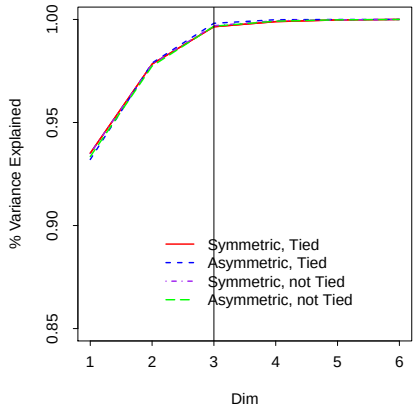


(b) Scenario 2

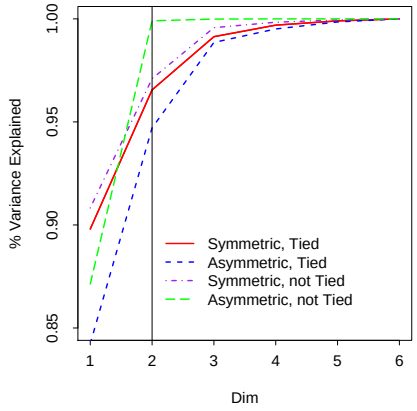


(c) Scenario 3

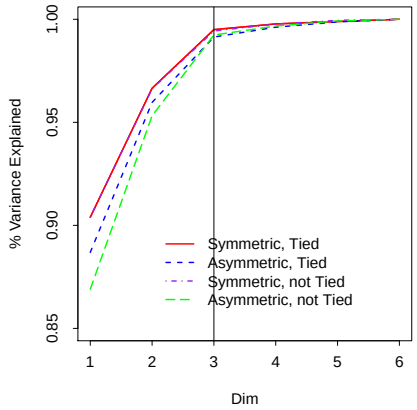
Figure 4.4: Deviance information criteria as a function of the embedding space's dimension K for simulated data sets. Vertical lines represent the underlying true latent dimension.



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

Figure 4.5: PNS decomposition of the latent space for the simulated data sets. Vertical lines represent the underlying true latent dimension.

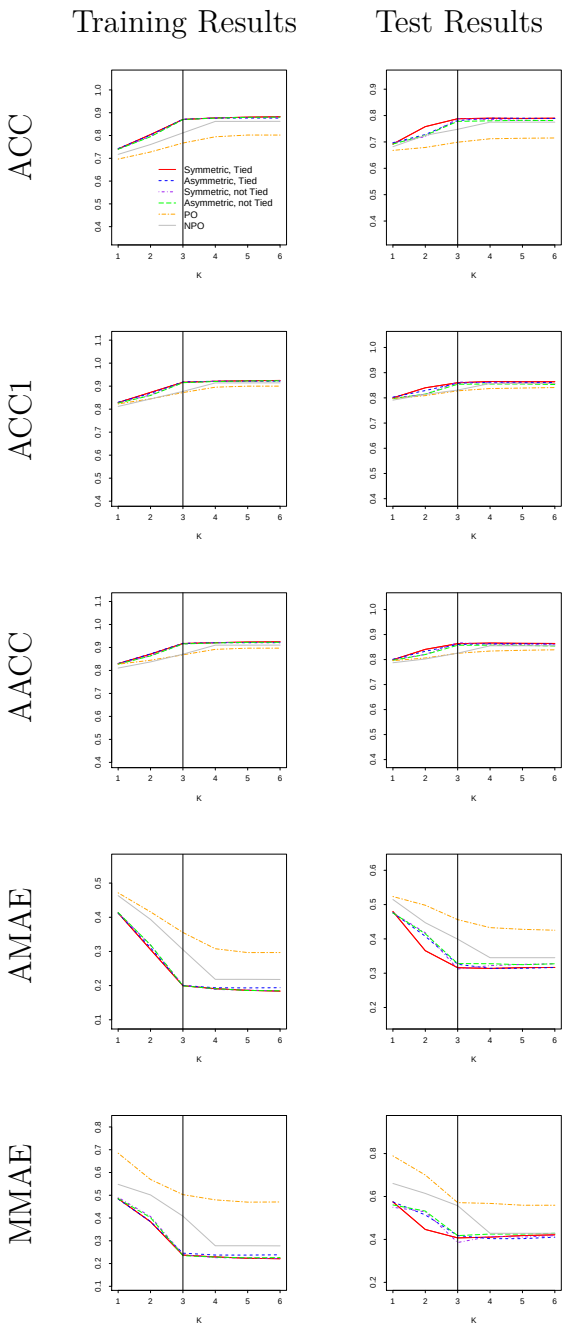


Figure 4.6: Nested model results for Scenario 1. Vertical lines represent the underlying true latent dimension.

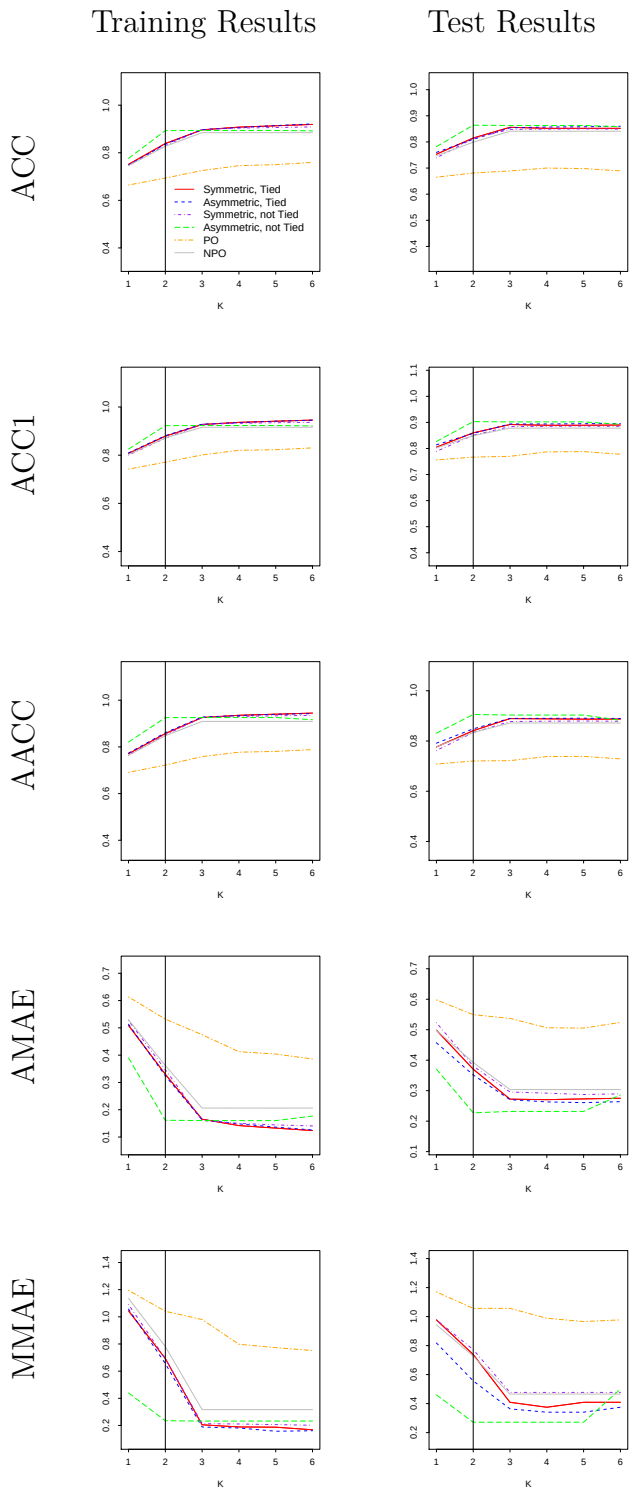


Figure 4.7: Nested model results for Scenario 2. Vertical lines represent the underlying true latent dimension.

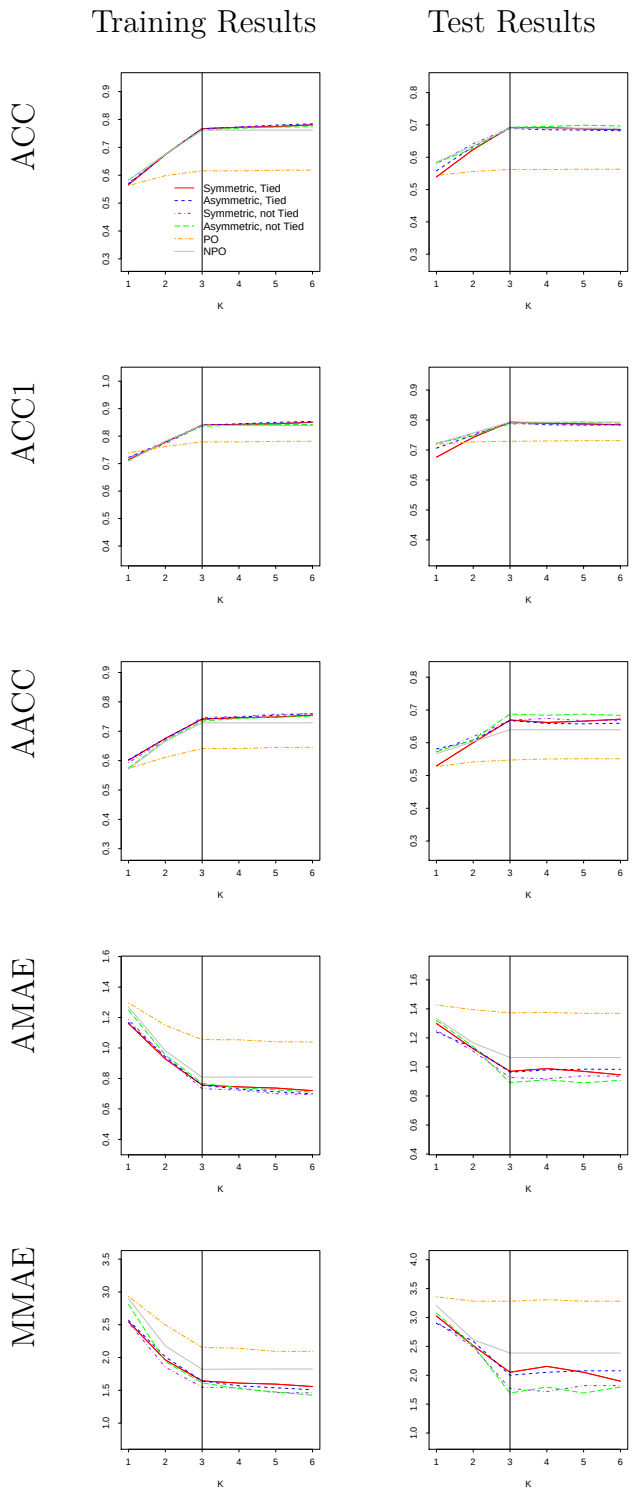


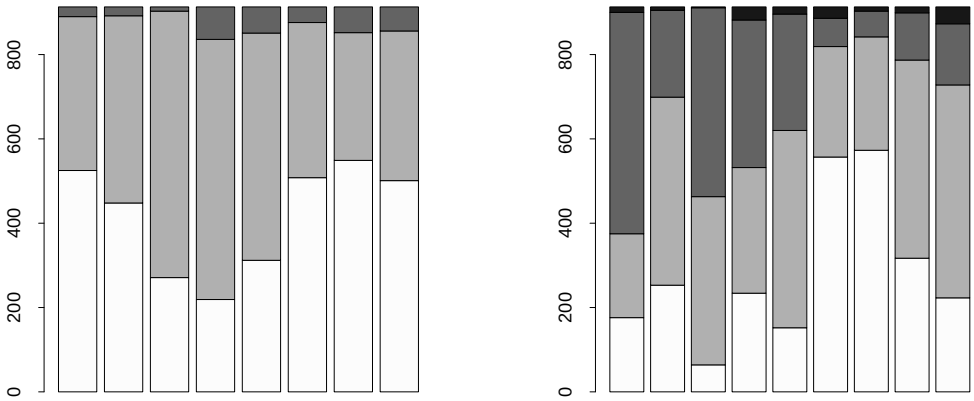
Figure 4.8: Nested model results for Scenario 3. Vertical lines represent the underlying true latent dimension.

4.5.2 Real data

In this Section, we illustrate the performance of the proposed models on two real data sets with unbalanced categories: the ASES and BEPS datasets. The ASES (Asia Europe Survey) data set [96] consists of responses by 913 interviewees and 17 questions focused on various political and economical issues in the U.K. Of these 17 questions, 9 share a four-level ordinal scale, while the remaining 8 share a three-level scale. The BEPS data set (available from https://www.britishelection study.com/data-object/2005-2009-bes-6-wave-panel-survey/) corresponds to British Election Study Six-Wave Panel Survey conducted through the period of 2005-2009. There were 7793 participants and we select 42 ordinal response questions concerning various social, political and economical issues to carry out our analysis. Of the 42 questions, 38 have a five-level categorical scale, three have a four-level scale, and one question has a three-level scale. Table 4.2 provides some summaries of these two data sets while Figure 4.9 and 4.10 illustrates the unbalanced nature of these data sets. ASES data does not contain any missing values while BEPS does. In this illustration we assume that the missing responses are missing completely at random.

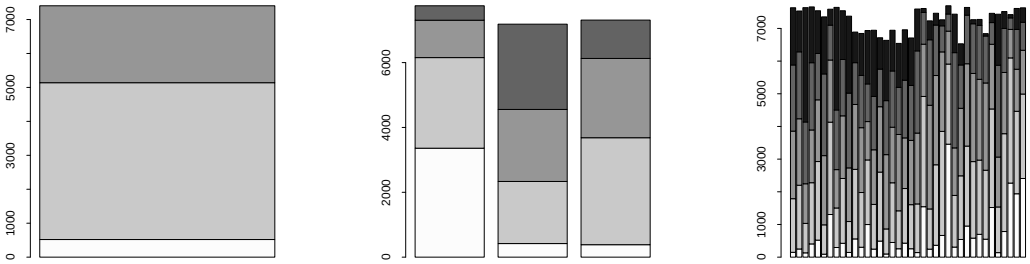
Data set	Surveyor (I)	Questions (J)	Missing Responses
ASES	913	17	0 (0.00%)
BEPS	7793	43	25431 (7.59%)

Table 4.2: Summary information for the two data sets analyzed in this chapter.



(a) Three-category **(b)** Four-category

Figure 4.9: Marginal distribution of the answers to each item in the ASES data, the left panel corresponds to three-category questions while the right panel corresponds to four-category questions. The ordinal scale is represented by the shades of gray from the lightest to the darkest.



(a) Three-category **(b)** Four-category **(c)** Five-category

Figure 4.10: Marginal distribution of the answers to each item in the BEPS data. The left panel corresponds to a three-category question, the middle panel corresponds to four-category questions, and the right panel corresponds to five-category questions. The ordinal scale is represented by the shades of gray from the lightest to the darkest (Note that the varying bar length are due to missing values).

Figures 4.15 and Figure 4.16 show the value of DIC for various models as a function of the latent dimension K. For the ASES data, DIC selects the two-dimensional NPO Euclidean model as the best model, followed by the one-dimensional NPO Euclidean model. Among the spherical models, DIC consistently selects one-dimensional models no matter what the exact form of the link function. Furthermore, among these spherical models, those with a symmetric link functions seem to be preferred. In contrast to the ASES dataset, for the BEPS data, DIC chooses a four-dimensional spherical model with tied and symmetric link function as the best model.

Figures 4.13 and 4.14 show the variance associated with each component of a PNS decomposition for the ASES and BEPS datasets, respectively. In the ASES dataset, with the exception of the spherical model with untied and asymmetric link function, the first dimension of the decomposition already accounts for 90% of the variance, while the first two account for over 96%. This suggests, that one or two latent dimensions are enough in this dataset, a result that agrees with those from Figure 4.15. On the other hand, we observe an elbow at dimension 4 in each spherical model for the BEPS data. Furthermore, the first three dimensions account for 93% of the variance and the first four dimensions account for more than 95% of the variance. Again, these behavior agrees with that observed in Figure 4.16 for DIC.

The nested and full model results for ASES data are shown in Figure 4.11. All models perform similarly in terms of the traditional measures. However, the PO Euclidean model outperforms the rest in the robust measures by a clear margin. On the other hand, the nested model results suggest a one-dimensional latent space for each model class as the higher-dimensional models do not provide a significant improvement in terms of the performance measures. This result agrees with the

DIC and PNS results discussed earlier. Next we discuss the nested and the full model results for the BEPS data shown in Figure 4.12. We observe that the the spherical and Euclidean models achieve similar results in the traditional measures but the spherical models excel at the robust measures. In addition, the results suggest roughly a four-dimensional latent space for each model by examining the elbow, which again are in accordance with the DIC and PNS results. In this case, the latent space clearly favors spherical latent space. Therefore, we consider the four-dimensional spherical model with the symmetric and tied link function as the best model from a simple parsimony argument.

Lastly, we conclude this section by providing examples for the item specific positions in which Euclidean and spherical latent space is favored respectively in Figure 4.17 and 4.18. Question 4 and 36 clearly exhibits spherical pattern while question 14 and 40 favors Euclidean. In particular, the item specific positions of question 4 and 36 are more scattered around the sphere while those of question 14 and 40 concentrates around the south pole with little variance. In addition, the black pair in question 4 and the blue pair in question 36 are positioned around opposing poles, which is similar to the circular bill examples discussed extensively in Section 2.4.1.

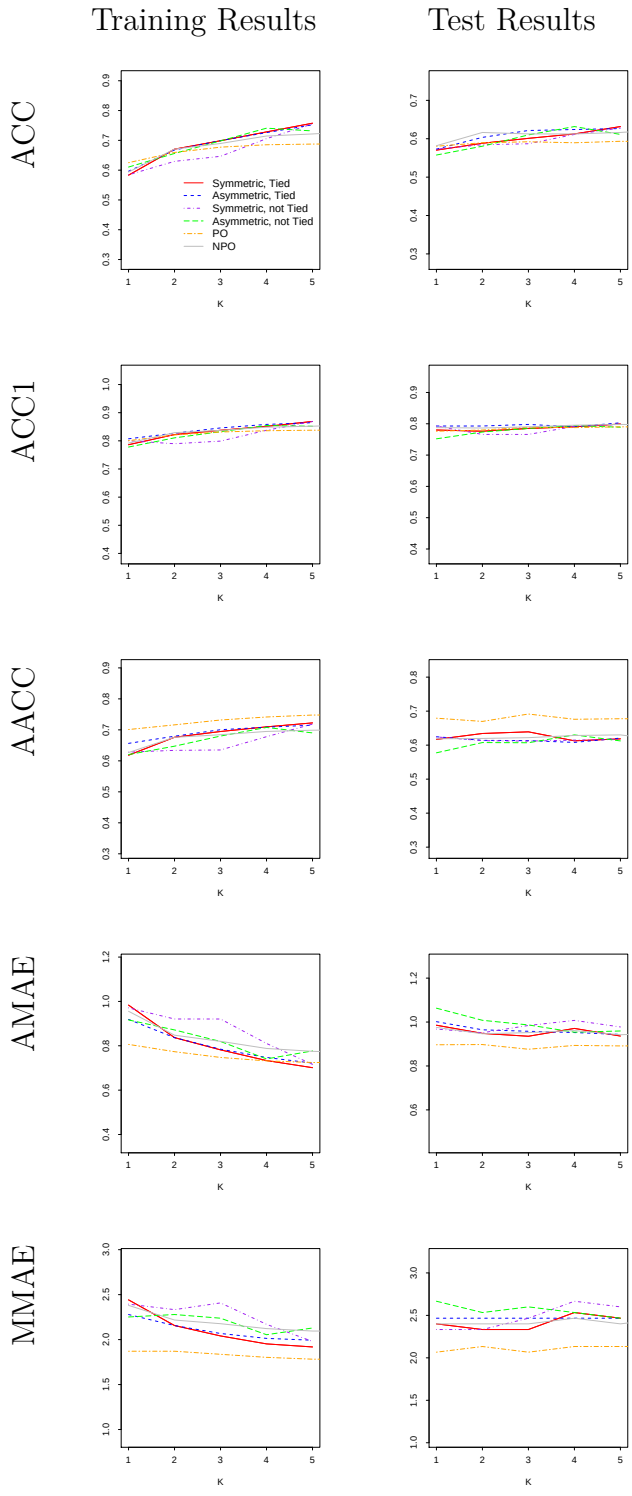


Figure 4.11: Nested model results of ASES data

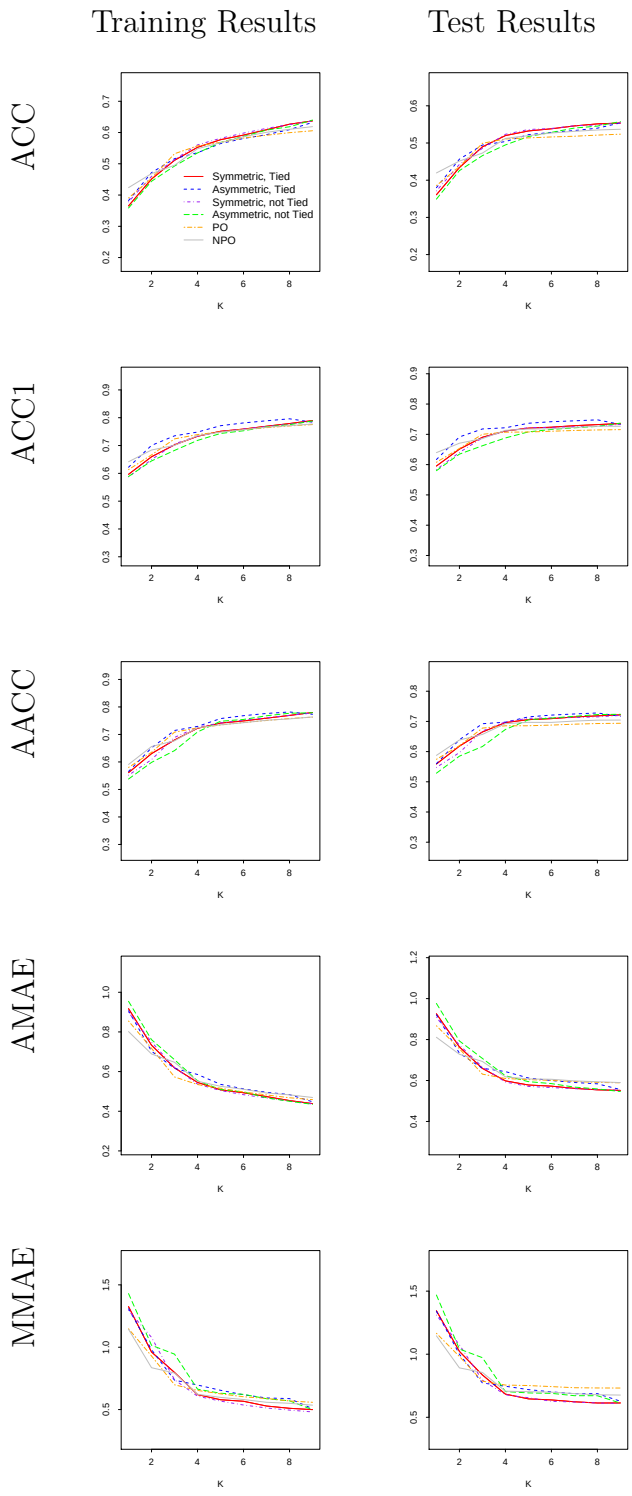


Figure 4.12: Nested model results for BEPS data

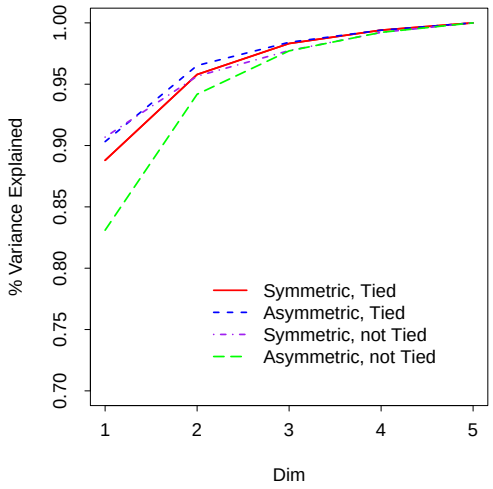


Figure 4.13: PNS decomposition of the latent space in the ASES data

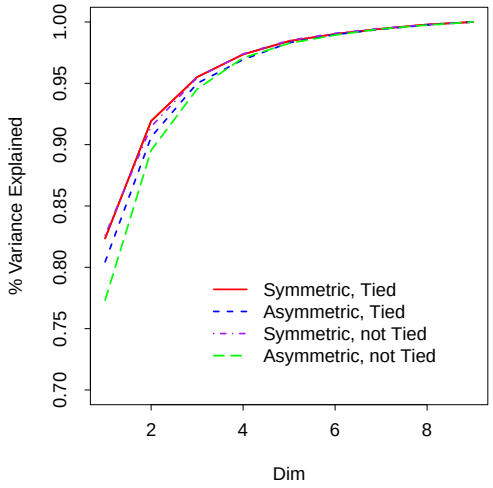


Figure 4.14: PNS decomposition of the latent space in the BEPS data

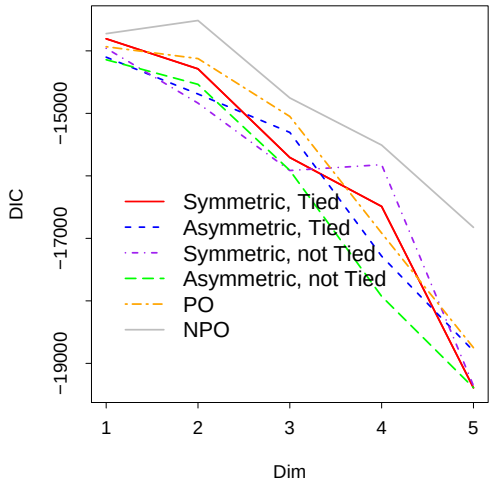


Figure 4.15: Deviance information criteria as a function of the embedding space's dimension K for ASES data

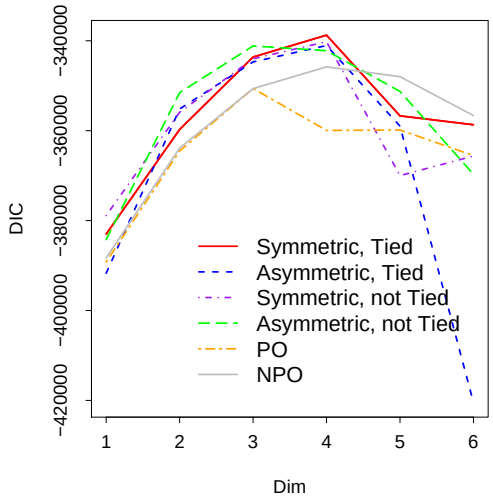
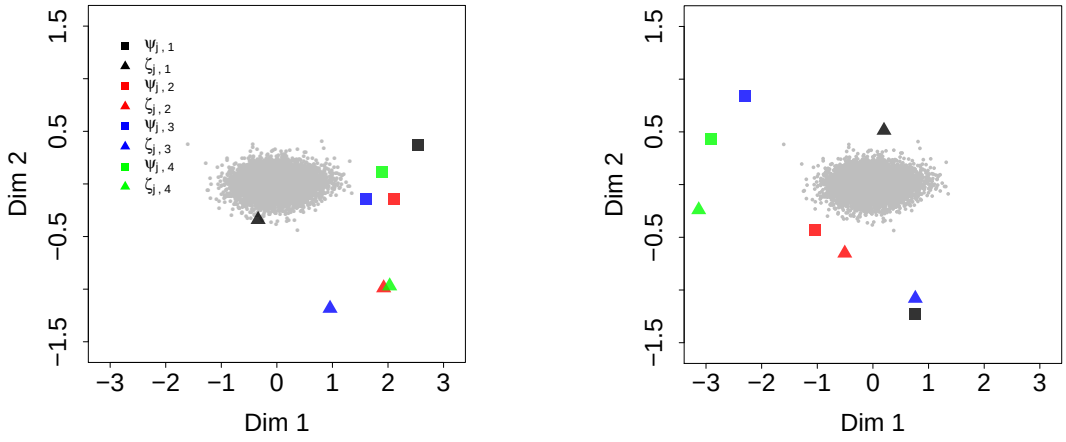
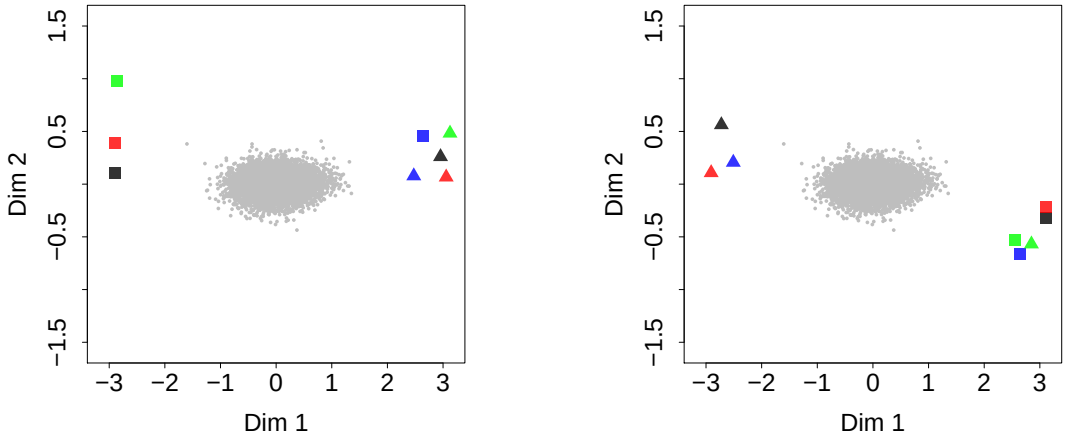


Figure 4.16: Deviance information criteria as a function of the embedding space's dimension K for BEPS data



(a) Question 4 **(b)** Question 36

Figure 4.17: Pairwise plots of the first two dimensions for two spherical questions. Points in gray represents the first two dimensions of $\boldsymbol{\beta}_i$s. Each pair of item specific position has the same color and the color represents different response category.



(a) Question 14 **(b)** Question 40

Figure 4.18: Pairwise plots of the first two dimensions for two Euclidean questions. Points in gray represents the first two dimensions of $\boldsymbol{\beta}_i$s. Each pair of item specific position has the same color and the color represents different response category.

4.6 Discussion

In this chapter, we demonstrated that our model is also capable of embedding ordinal data into the latent space through careful selection of the ordinal structure as well as various configurations of the link function. We evaluate our model using both simulated and real datasets based on both traditional and robust performance measures. We conclude that our spherical model can closely approximate traditional factor models when the true latent space is Euclidean, but yield superior results when the latent space is indeed spherical. In addition, our results suggest that we can employ the variance decomposition method and the nested model results computed from the full model to select the optimal latent dimension as well as perform model comparison without fitting models of varying dimensions.

Chapter 5

Conclusion and Future Works

In this thesis, we proposed a general framework of embedding binary and categorical data into the spherical latent space. We demonstrated that it is possible to distinguish between dimensions and geometries of the underlying embedding space. In addition, our model can approximate the traditional Euclidean factor model closely when the latent space is indeed Euclidean. Furthermore, we justified the use of such space theoretically and practically through both simulated and real data sets. While we acknowledge that a spherical latent space might not be appropriate in every application (e.g., it would be difficult to justify in application such as educational testing), we believe that the class of models discussed in this dissertation can have broader applications beyond the analysis of roll call data. One such area of broader application is marketing, where choice models are widely used to understand consumer behavior. In this context, spherical models could serve to explain the apparent lack of transitivity of preferences that is sometimes present in real marketing data (e.g., see [97]).

Another potential line of research is to incorporate covariates into the spherical

model. In the context of roll call voting application, legislator's personal characteristics and backgrounds such as party and committee membership, gender, ethnic background, age, education background, the type of constituency represented (eg., rural, urban, agricultural) may very well shape their policy decision. Therefore, incorporating this information will likely lead to a better representation of legislator's latent policy space as well as helping researchers to learn the factors that shape their legislative decision making. To accomplish this, we could adapt method discussed in 98, pg. 258 to incorporate covariates into the one-dimensional spherical model through the prior associated with each legislator. In particular, we could let

$$\beta_i \mid \omega_\beta, \tau_\beta, \mathbf{h} \sim \text{vonMis}(\tau_\beta + g(\mathbf{h}^T \boldsymbol{x}_i), \omega_\beta), \tag{5.1}$$

where $\boldsymbol{x}_i$s are covariates, $\mathbf{h}$ represents the regression coefficients and $g(x)$ is a one-to-one link function which maps the real line onto $(-\pi, \pi)$. One such link function $g(x)$ is $2\tan^{-1}(x)$. This approach can be further extended to higher dimensional latent spaces in a straightforward manner, resulting in a spherical generalization of structural equation models.

We relied on the DIC to perform model comparison in Chapter 2 and 3. In future work, we will investigate the use of shrinkage prior such as spike and slab prior [99] or the gamma process shrinkage prior [100] as alternatives to DIC for dimensionality selection. We note, however, that a direct implementation may be challenging because there is a fundamental difference between the Euclidean and spherical manifold. More specifically, that for dimension k to be inactive for observation $y_{i,j}$, we need $\beta_{i,k} = \psi_{j,k} = \zeta_{j,k} = 0$ (recall Section 3.1.1). Ensuring this requires a joint prior specification for all three parameters. This is different from what happens with the Euclidean model where having $\beta_{i,k} = 0$ is enough for

the k-th dimension to be inactive.

While the focus of this thesis has been on spherical latent spaces, it is clear that our approach can be extended to other classes of embedding manifolds such as affine subspaces, Stiefel manifolds or product manifolds as long as the associated geodesic is explicit and known. The main challenge associated with this kind of extension is also the construction of prior distributions as illustrated in Chapter 3, specially if we aim to estimate the intrinsic dimension of the space. The discussion in Section 3.4 surrounding the use of alternative prior distributions for the latent positions that concentrate their mass on small spheres with radius less than one (such as those in [86] and [87]) can also be seen as part of these future efforts.

Similarly, this class of models can be extended beyond binary and ordinal to nominal observations using similar random utility formulations. For example, an immediate extension of our model that embeds nominal data can be achieved by adapting the NPO adjacent-category structure discussed in Section 4.1.2 into our spherical model framework.

Appendix A

Appendix

A.1 Hausdorff measures and their gradients for the circular factor model

The density of the full conditional distribution for β_i is given by

$$
\begin{aligned}
p(\beta_i \mid \cdots) \propto & \exp\{\omega_\beta \beta_i\} \\
& \prod_{j=1}^{J} \left[G_{\kappa_j} \left(\{\arccos(\cos(\zeta_j - \beta_i))\}^2 - \{\arccos(\cos(\psi_j - \beta_i))\}^2 \right) \right]^{y_{i,j}} \\
& \qquad \left[1 - G_{\kappa_j} \left(\{\arccos(\cos(\zeta_j - \beta_i))\}^2 - \{\arccos(\cos(\psi_j - \beta_i))\}^2 \right) \right]^{1-y_{i,j}}
\end{aligned}
$$

Then, the density of the associated Hausdorff measure in $\mathbb{R}^2$ is given by

$$
\begin{aligned}
p(\boldsymbol{x}_{\beta_i} \mid \cdots) \propto & \exp\left\{\boldsymbol{\eta}_\beta^T \boldsymbol{x}_{\beta_i}\right\} \\
& \prod_{j=1}^{J}\left[G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right)\right]^{y_{i,j}} \\
& \left[1 - G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right)\right]^{1-y_{i,j}}, \ \boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1,
\end{aligned}
$$

where $\boldsymbol{\eta}_\beta^T = (\omega_\beta, 0)$, $\boldsymbol{z}_{\psi_j}^T = (\cos\psi_j, \sin\psi_j)$, $\boldsymbol{z}_{\zeta_j}^T = (\cos\zeta_j, \sin\zeta_j)$, and the mapping between β_i and $\boldsymbol{x}_{\beta_i}$ is given by $\boldsymbol{x}_{\beta_i}^T = (\cos\beta_i, \sin\beta_i)$. Hence, the gradient of the Hausdorff measure is simply

$$
\nabla \log_{\mathcal{H}} p(\boldsymbol{x}_{\beta_i} \mid \cdots) = \boldsymbol{\eta}_\beta + \sum_{j=1}^{J}\begin{pmatrix} y_{i,j} e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \\ y_{i,j} e'_{i,j,2} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,2} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \end{pmatrix},
$$

where

$$
e_{i,j} = \{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2,
$$

$$
\boldsymbol{e}'_{i,j} = \begin{pmatrix} e'_{i,j,1} \\ e'_{i,j,2} \end{pmatrix} = \frac{2\arccos\left(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1-\left(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \boldsymbol{z}_{\psi_j} - \frac{2\arccos\left(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1-\left(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \boldsymbol{z}_{\zeta_j}.
$$

Next, the density of the full conditional distribution for $\boldsymbol{z}_{\zeta_j}$ in terms of Hausdorff measure in $\mathbb{R}^2$ is given by

$$
\begin{aligned}
p(\boldsymbol{z}_{\zeta_j} \mid \cdots) \propto & \prod_{i=1}^{I} G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right) \\
& \left[1 - G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right)\right]^{1-y_{i,j}}, \ \boldsymbol{z}_{\zeta_j}^T \boldsymbol{z}_{\zeta_j} = 1,
\end{aligned}
$$

and the gradient of the Hausdorff measure is

$$\nabla \log_{\mathcal{H}} p(\boldsymbol{z}_{\zeta_j} \mid \cdots) = \sum_{i=1}^{I} \begin{pmatrix} y_{i,j}\delta'_{i,j,1}\frac{g_{\kappa_j}(\delta_{i,j})}{G_{\kappa_j}(\delta_{i,j})} - (1-y_{i,j})\delta'_{i,j,1}\frac{g_{\kappa_j}(\delta_{i,j})}{1-G_{\kappa_j}(\delta_{i,j})} \\ y_{i,j}\delta'_{i,j,2}\frac{g_{\kappa_j}(\delta_{i,j})}{G_{\kappa_j}(\delta_{i,j})} - (1-y_{i,j})\delta'_{i,j,2}\frac{g_{\kappa_j}(\delta_{i,j})}{1-G_{\kappa_j}(\delta_{i,j})} \end{pmatrix},$$

where $\delta_{i,j} = \{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2$,

$$\boldsymbol{\delta}'_{i,j} = \begin{pmatrix} \delta'_{i,j,1} \\ \delta'_{i,j,2} \end{pmatrix} = -\frac{2\arccos\left(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1-\left(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \boldsymbol{x}_{\beta_i}.$$

Lastly, the density of the full conditional distribution for $\boldsymbol{z}_{\psi_j}$ in terms of Hausdorff measure in $\mathbb{R}^2$ is given by

$$p(\boldsymbol{z}_{\psi_j} \mid \cdots) \propto \prod_{i=1}^{I} G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right)$$
$$\left[1 - G_{\kappa_j}\left(\{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2\right)\right]^{1-y_{i,j}}, \ \boldsymbol{z}_{\psi_j}^T \boldsymbol{z}_{\psi_j} = 1,$$

and the gradient of the Hausdorff measure is

$$\nabla \log_{\mathcal{H}} p(\boldsymbol{z}_{\psi_j} \mid \cdots) = \sum_{i=1}^{I} \begin{pmatrix} y_{i,j}\gamma'_{i,j,1}\frac{g_{\kappa_j}(\gamma_{i,j})}{G_{\kappa_j}(\gamma_{i,j})} - (1-y_{i,j})\gamma'_{i,j,1}\frac{g_{\kappa_j}(\delta_{i,j})}{1-G_{\kappa_j}(\gamma_{i,j})} \\ y_{i,j}\gamma'_{i,j,2}\frac{g_{\kappa_j}(\gamma_{i,j})}{G_{\kappa_j}(\gamma_{i,j})} - (1-y_{i,j})\gamma'_{i,j,2}\frac{g_{\kappa_j}(\delta_{i,j})}{1-G_{\kappa_j}(\gamma_{i,j})} \end{pmatrix},$$

where

$$\gamma_{i,j} = \{\arccos(\boldsymbol{z}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2,$$

$$\boldsymbol{\gamma}'_{i,j} = \begin{pmatrix} \gamma'_{i,j,1} \\ \gamma'_{i,j,2} \end{pmatrix} = \frac{2\arccos\left(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1-\left(\boldsymbol{z}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \boldsymbol{x}_{\beta_i}$$

A.2 Stability of priors in the Euclidean factor model

Let $\mu_j \sim N(0, 1/2)$, $\alpha_{j,k} \sim N(0, 1/2)$, $\beta_{j,k} \sim N(0, 6/[\pi k]^2)$, and $z_{i,j}(K) = \mu_j + \sum_{k=1}^{K} \alpha_{j,k}\beta_{i,k}$. Because $z_{i,j}(K)$ is the sum of $K+1$ random variables with finite second moments, a simple application of the central limit theorem indicates that $\{z_{i,j}(K) - \mathrm{E}(z_{i,j}(K))\}/\mathrm{Var}(z_{i,j}(K))$ converges in distribution to standard normal distribution as $K \to \infty$. Now, note that $(z_{i,j}(K)) = 0$ by construction, and that,

$$\begin{aligned}\mathrm{Var}\left\{z_{i,j}(K)\right\} &= \mathrm{Var}(\mu_j) + \sum_{k=1}^{K} \mathrm{Var}(\alpha_{j,k})\mathrm{Var}(\beta_{i,k}) \\ &= \frac{1}{2} + \frac{1}{2}\sum_{k=1}^{K} \frac{6}{\pi^2}\frac{1}{k^2} = \frac{1}{2} + \frac{1}{2}\frac{6}{\pi^2}\frac{\pi^2}{6} = 1\end{aligned}$$

As a consequence, $\theta_{i,j} = \Phi\left(z_{i,j}(K)\right)$ converges in distribution to a uniform distribution in $[0, 1]$.

A.3 Hausdorff measure of SvM distribution

$$\begin{aligned}p(\boldsymbol{\phi} \mid \boldsymbol{\omega}) &= \left(\frac{1}{2\pi}\right)^K 2^{K-1}\frac{1}{I_0(\omega_1)}\exp\left\{\omega_1\cos\phi_1\right\}\prod_{k=2}^{K}\frac{1}{I_0(\omega_k)}\exp\left\{\omega_k\cos 2\phi_k\right\} \\ &= \frac{1}{2\pi I_0(\omega_1)}\exp\left\{\omega_1\cos\phi_1\right\}\left\{\prod_{k=2}^{K}\frac{1}{\pi I_o(\omega_k)}\exp\left(\omega_k\cos 2\phi_k\right)\right\} \\ &= \frac{1}{2\pi I_0(\omega_1)}\exp\left\{\omega_1\cos\phi_1\right\}\left\{\prod_{k=2}^{K}\frac{1}{\pi I_o(\omega_k)}\exp\left(\omega_k(2\cos^2\phi_k - 1)\right)\right\} \\ &= \frac{1}{2\pi I_0(\omega_1)}\exp\left\{\omega_1\cos\phi_1\right\}\left\{\prod_{k=2}^{K}\frac{1}{\pi I_o(\omega_k)}\right\}\exp\left\{\sum_{k=2}^{K}\omega_k\left(2\cos^2\phi_k - 1\right)\right\}.\end{aligned}$$

Using the hyperspherical coordinates in Equation 3.3, we can express our prior

proposed prior in terms of Hausdorff measure,

$$p(\boldsymbol{x} \mid \boldsymbol{\omega}) = |\boldsymbol{J}| \times \ p\left(\boldsymbol{\phi} = g(\boldsymbol{x}) \mid \boldsymbol{\omega}\right) = \left\{ \frac{1}{\prod_{k=1}^{K} \sqrt{\sum_{t=1}^{k+1} x_t^2}} \right\} \frac{1}{2\pi I_o(\omega_1)}$$
$$\exp\left\{ \omega_1 \frac{x_1}{\sqrt{x_1^2 + x_2^2}} \right\} \left\{ \prod_{k=2}^{K} \frac{1}{\pi I_o(\omega_k)} \right\} \exp\left\{ -\sum_{k=2}^{K} \omega_k \left(2\frac{x_{k+1}^2}{\sum_{t=1}^{k+1} x_t^2} - 1 \right) \right\}, \quad \boldsymbol{x}^T \boldsymbol{x} = 1.$$

A.4 Hausdorff measure and their gradients of the spherical latent factor model for binary data

A.4.1 Gradients of the loglikelihood

The gradients with respect to the Hausdorff measure for $\boldsymbol{x}_{\beta_i}, \boldsymbol{x}_{\zeta_j}, \boldsymbol{x}_{\psi_j}$ are derived under the unit norm constraints in which their first K components are independent variables while the last dimension is dependent. Therefore the gradient for the last dimension is always 0 and the gradients shown in this Section are for the first K dimension.

The gradient of log conditional density of $\boldsymbol{x}_{\beta_i}$ under the Hausdorff measure is,

$$\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \dots) = \nabla \log p_{\mathcal{H}_l} + \nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J,$$

where $\nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J$ is defined in (A.2). The density of the associated

Hausdorff measure with respect to the likelihood component is given by

$$p(\boldsymbol{x}_{\beta_i} \mid \cdots) = \prod_{j=1}^{J} \left[G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{y_{i,j}}$$
$$\left[1 - G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{1-y_{i,j}}, \quad \boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1,$$

where $\boldsymbol{x}_{\beta_i} = (x_{\beta_i,1}, x_{\beta_i,2}, \ldots, x_{\beta_i,K+1})$, $\boldsymbol{x}_{\zeta_j} = \left(x_{\zeta_j,1}, x_{\zeta_j,2}, \ldots, x_{\zeta_j,K+1}\right)$, $\boldsymbol{x}_{\psi_j} = \left(x_{\psi_j,1}, x_{\psi_j,2}, \ldots, x_{\psi_j,K+1}\right)$ Hence, the gradient of the likelihood under the Hausdorff measure is simply

$$\nabla \log_{\mathcal{H}_l} p(\boldsymbol{x}_{\beta_i} \mid \cdots) = \begin{pmatrix} \sum_{j=1}^{J} \left\{ y_{i,j} e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \right\} \\ \vdots \\ \sum_{j=1}^{J} \left\{ y_{i,j} e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \right\} \end{pmatrix},$$

where $e_{i,j} = \{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2$,

$$e'_{i,j,t} = \frac{2 \arccos\left(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1 - \left(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \left(x_{\psi_{j,t}} - x_{\psi_{j,K+1}} \frac{x_{\beta_{i,t}}}{x_{\beta_{i,K+1}}} \right) -$$
$$\frac{2 \arccos\left(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1 - \left(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \left(x_{\zeta_{j,t}} - x_{\zeta_{j,K+1}} \frac{x_{\beta_{i,t}}}{x_{\beta_{i,K+1}}} \right).$$

Next the gradient of log conditional density of $\boldsymbol{x}_{\zeta_j}$ under the Hausdorff measure is,

$$\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\zeta_j} \mid \ldots) = \nabla \log p_{\mathcal{H}_l} + \nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J,$$

where $\nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J$ is defined in (A.2). The density of the associated

Hausdorff measure with respect to the likelihood component is given by

$$p(\boldsymbol{x}_{\zeta_j} \mid \cdots) = \prod_{i=1}^{I} \left[G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{y_{i,j}}$$
$$\left[1 - G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{1-y_{i,j}}, \quad \boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\zeta_j} = 1,$$

where $\boldsymbol{x}_{\beta_i} = (x_{\beta_i,1}, x_{\beta_i,2}, \ldots, x_{\beta_i,K+1})$, $\boldsymbol{x}_{\zeta_j} = \left(x_{\zeta_j,1}, x_{\zeta_j,2}, \ldots, x_{\zeta_j,K+1}\right)$, $\boldsymbol{x}_{\psi_j} = \left(x_{\psi_j,1}, x_{\psi_j,2}, \ldots, x_{\psi_j,K+1}\right)$ Hence, the gradient of the likelihood under the Hausdorff measure is simply

$$\nabla \log_{\mathcal{H}_l} p(\boldsymbol{x}_{\zeta_j} \mid \cdots) = \begin{pmatrix} \sum_{i=1}^{I} \left\{ y_{i,j} e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1 - y_{i,j}) e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{1 - G_{\kappa_j}(e_{i,j})} \right\} \\ \vdots \\ \sum_{i=1}^{I} \left\{ y_{i,j} e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1 - y_{i,j}) e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{1 - G_{\kappa_j}(e_{i,j})} \right\} \end{pmatrix},$$

where

$$e_{i,j} = \{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2,$$

$$e'_{i,j,t} = -\frac{2 \arccos\left(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1 - \left(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \left(x_{\beta_{i,t}} - x_{\beta_{i,K+1}} \frac{x_{\zeta_{j,t}}}{x_{\zeta_{j,K+1}}} \right).$$

Lastly the gradient of log conditional density of $\boldsymbol{x}_{\psi_j}$ under the Hausdorff measure is,

$$\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\psi_j} \mid \ldots) = \nabla \log p_{\mathcal{H}_l} + \nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J,$$

where $\nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J$ is defined in (A.2). The density of the associated

Hausdorff measure with respect to the likelihood component is given by

$$p(\boldsymbol{x}_{\zeta_j} \mid \cdots) = \prod_{i=1}^{I} \left[G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{y_{i,j}}$$
$$\left[1 - G_{\kappa_j} \left(\{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 \right) \right]^{1-y_{i,j}}, \quad \boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\psi_j} = 1,$$

where $\boldsymbol{x}_{\beta_i} = (x_{\beta_i,1}, x_{\beta_i,2}, \ldots, x_{\beta_i,K+1})$, $\boldsymbol{x}_{\zeta_j} = \left(x_{\zeta_j,1}, x_{\zeta_j,2}, \ldots, x_{\zeta_j,K+1}\right)$, $\boldsymbol{x}_{\psi_j} = \left(x_{\psi_j,1}, x_{\psi_j,2}, \ldots, x_{\psi_j,K+1}\right)$ Hence, the gradient of the likelihood under the Hausdorff measure is simply

$$\nabla \log_{\mathcal{H}_l} p(\boldsymbol{x}_{\psi_j} \mid \cdots) = \begin{pmatrix} \sum_{i=1}^{I} \left\{ y_{i,j} e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,1} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \right\} \\ \vdots \\ \sum_{i=1}^{I} \left\{ y_{i,j} e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{G_{\kappa_j}(e_{i,j})} - (1-y_{i,j}) e'_{i,j,K} \frac{g_{\kappa_j}(e_{i,j})}{1-G_{\kappa_j}(e_{i,j})} \right\} \end{pmatrix},$$

where

$$e_{i,j} = \{\arccos(\boldsymbol{x}_{\zeta_j}^T \boldsymbol{x}_{\beta_i})\}^2 - \{\arccos(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i})\}^2,$$

$$e'_{i,j,t} = \frac{2 \arccos\left(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)}{\sqrt{1 - \left(\boldsymbol{x}_{\psi_j}^T \boldsymbol{x}_{\beta_i}\right)^2}} \left(x_{\beta_{i,t}} - x_{\beta_{i,K+1}} \frac{x_{\psi_{j,t}}}{x_{\psi_{j,K+1}}} \right).$$

A.4.2 Derivation of the gradient of log prior and log Jacobian

The gradient with respect to the log of Hausdorff measure may appear forbidding to derive, especially the prior component of the full conditional because the expression of the gradient changes with the predetermined maximum dimension. Hence, we develop an automatic method to facilitate the computation of the gradient. Interestingly, the GHMC algorithm projects all the gradient to the tangent space through the momentum update, which allows the gradient computed with or

without the constraint to be the same. Nevertheless, it is generally recommended to derive the gradients under the unit norm constraint since they are invariant to the reparameterization with respect to the constraint.

$$\nabla \log p_{\mathcal{H}}(\boldsymbol{x}_{\beta_i} \mid \dots) = \nabla \log p_{\mathcal{H}_l} + \nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_J, \tag{A.1}$$

where $\log p_{\mathcal{H}_l}$, $\log p_{\mathcal{H}_\pi}$ and $\nabla \log p_J$ is the gradient of log-likelihood, log of prior and log of Jacobian respectively with respect to the Hausdorff measure. For $K = 1$, the model becomes the circular factor model and we focus on the derivation for $K > 1$ in this paper. Let $x_{\beta_{i,1}}, \dots, x_{\beta_{i,K}}$ be the independent random variable and $\boldsymbol{x}_{\beta_{i,K+1}}$ be the dependent random variable. As a result, the gradient associated with the last dimension $x_{\beta_{i,K+1}}$ is 0. The first K components of the gradient from the prior and Jacobian are shown as follows,

$$\begin{aligned}
\nabla \log p_{\mathcal{H}_\pi} + \nabla \log p_{\mathcal{H}_J} &= 4\omega_K \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} + \omega_1 x_2 (x_1^2 + x_2^2)^{-\frac{3}{2}} \begin{bmatrix} x_2 \\ -x_1 \\ 0 \\ \vdots \end{bmatrix} \\
&+ 4 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} \circ \Sigma_{\text{-}2} \begin{bmatrix} \omega_2 x_3^2 / \left(\sum_{t=1}^{3} x_t^2 \right)^2 \\ \vdots \\ \omega_{K-2} x_{K-1}^2 / \left(\sum_{t=1}^{K-1} x_t^2 \right)^2 \\ \omega_{K-1} x_K^2 / \left(\sum_{t=1}^{K} x_t^2 \right)^2 \end{bmatrix} - 4 \begin{bmatrix} 0 \\ 0 \\ \frac{\omega_2 x_3}{\sum_{t=1}^{3} x_t^2} \\ \vdots \\ \frac{\omega_{K-1} x_K}{\sum_{t=1}^{K} x_t^2} \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix} \circ \Sigma_{\text{-}1} \begin{bmatrix} \frac{1}{\sum_{t=1}^{2} x_t^2} \\ \frac{1}{\sum_{t=1}^{3} x_t^2} \\ \vdots \\ \frac{1}{\sum_{t=1}^{K} x_t^2} \end{bmatrix},
\end{aligned} \tag{A.2}$$

where Σ_{-t} is an upper triangular matrix of size $K-t$ without the first t columns and all non-zero entries are 1. As a special case when $K=2$, Σ_{-2} becomes a 0 matrix and Σ_{-1} becomes scalar 1. Once the maximum dimension is specified, Σ_{-1}, Σ_{-2} can be generated and all the gradients can then be vectored easily during the implementation without explicitly deriving the expression for different maximum dimension. Recall that, $\boldsymbol{x}_{\beta_i}, \boldsymbol{x}_{\zeta_j}$ and $\boldsymbol{x}_{\psi_j}$ share the same prior and hence the corresponding gradient expression will also be the same.

A.5 Hausdorff measure and their gradients of the spherical latent factor model for ordinal data

As discussed in Section 4.2, we employ the same spherical von Mises prior for $\boldsymbol{\psi}_{j,l}$s, $\boldsymbol{\zeta}_{j,l}$s and $\boldsymbol{\beta}_i$s of which the gradient is identical to those in Appendix A.5. Therefore, we focus the derivation of the likelihood ($\nabla \log p_{\mathcal{H}_l}$ in Equation (A.1) for $\boldsymbol{x}_{\beta_i}, \boldsymbol{x}_{\zeta_j}$ and $\boldsymbol{x}_{\psi_j}$.

The density of the associated Hausdorff measure with respect to the likelihood of $\boldsymbol{x}_{\beta_i}$ is given by,

$$p_{\mathcal{H}_l}(\boldsymbol{x}_{\beta_i} \mid \cdots) = \left[\prod_{j=1}^{J} \prod_{l=1}^{L_j} (\theta_{i,j}^{l} (\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}))^{z_{i,j}^{l}}, \right], \quad \boldsymbol{x}_{\beta_i}^T \boldsymbol{x}_{\beta_i} = 1,$$

and the corresponding gradient of the log density is as follows,

$$\nabla \log p_{\mathcal{H}_l}(\boldsymbol{x}_{\beta_i} \mid \cdots) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} z_{i,j}^{l} \log(\theta_{i,j}^{l} (\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l})),$$

where $z_{i,j}^l$ is an indicator variable and is 1 only if $Y_{i,j} = l$ and recall $\theta_{i,j}^l$ is defined Equation (4.21).

The density of the associated Hausdorff measure with respect to the likelihood of $\boldsymbol{x}_{\zeta_{j,l}}$ is given by,

$$p_{\mathcal{H}_l}(\boldsymbol{x}_{\zeta_{j,l}} \mid \cdots) = \left[\prod_{i=1}^{I} \prod_{l=1}^{L_j} (\theta_{i,j}^l \left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}\right))^{z_{i,j}^l}, \right], \quad \boldsymbol{x}_{\zeta_{j,l}}^T \boldsymbol{x}_{\zeta_{j,l}} = 1,$$

and the corresponding gradient of the log density is as follows,

$$\nabla \log p_{\mathcal{H}_l}(\boldsymbol{x}_{\zeta_{j,l}} \mid \cdots) = \sum_{i=1}^{I} \sum_{l=1}^{L_j} z_{i,j}^l \log(\theta_{i,j}^l \left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}\right)),$$

where $z_{i,j}^l$ is an indicator variable and is 1 only if $Y_{i,j} = l$ and recall $\theta_{i,j}^l$ is defined Equation (4.21).

The density of the associated Hausdorff measure with respect to the likelihood of $\boldsymbol{x}_{\psi_{j,l}}$ is given by,

$$p_{\mathcal{H}_l}(\boldsymbol{x}_{\psi_{j,l}} \mid \cdots) = \left[\prod_{i=1}^{I} \prod_{l=1}^{L_j} (\theta_{i,j}^l \left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}\right))^{z_{i,j}^l}, \right], \quad \boldsymbol{x}_{\psi_{j,l}}^T \boldsymbol{x}_{\psi_{j,l}} = 1,$$

and the corresponding gradient of the log density is as follows,

$$\nabla \log p_{\mathcal{H}_l}(\boldsymbol{x}_{\psi_{j,l}} \mid \cdots) = \sum_{i=1}^{I} \sum_{l=1}^{L_j} z_{i,j}^l \log(\theta_{i,j}^l \left(\boldsymbol{\beta}_i, \boldsymbol{\zeta}_{j,l}, \boldsymbol{\psi}_{j,l}\right)),$$

where $z_{i,j}^l$ is an indicator variable and is 1 only if $Y_{i,j} = l$ and recall $\theta_{i,j}^l$ is defined Equation (4.21).

Bibliography

[1] D. Child, *The essentials of factor analysis*. Bloomsbury Academic, 3 ed., 2006.

[2] S. A. Mulaik, *Foundations of factor analysis*. CRC press, 2 ed., 2009.

[3] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[4] J. Xu and G. Durrett, "Spherical latent spaces for stable variational autoencoders," *arXiv preprint arXiv:1808.10805*, 2018.

[5] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," *arXiv preprint arXiv:1804.00891*, 2018.

[6] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, pp. 585–591, 2002.

[10] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, pp. 313–338, 2004.

[11] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *Journal of machine learning research*, vol. 6, no. Nov, pp. 1783–1816, 2005.

[12] J. Marschak, "Binary-choice constraints and random utility indicators," in *Stanford Symposium on Mathematical Methods in the Social Sciences*, (Stanford, CA), Stanford University Press, 1960.

[13] D. McFadden, "Conditional logit analysis of qualitative choice behavior," in *Frontiers of Economics* (P. Zarembka, ed.), pp. 105–142, Institute of Urban and Regional Development, University of California, 1973.

[14] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE transactions on medical imaging*, vol. 23, no. 8, pp. 995–1005, 2004.

[15] S. Sommer, F. Lauze, and M. Nielsen, "The differential of the exponential map, jacobi fields and exact principal geodesic analysis," *CoRR, abs/1008.1902*, 2010.

[16] M. Zhang and T. Fletcher, "Probabilistic principal geodesic analysis," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 1178–1186, Curran Associates, Inc., 2013.

[17] S. Jung, I. L. Dryden, and J. Marron, "Analysis of principal nested spheres," *Biometrika*, vol. 99, no. 3, pp. 551–568, 2012.

[18] O. A. Davis, M. J. Hinich, and P. C. Ordeshook, "An expository development of a mathematical model of the electoral process," *American Political Science Review*, vol. 64, no. 2, pp. 426–448, 1970.

[19] J. M. Enelow and M. J. Hinich, *The spatial theory of voting: An introduction*. CUP Archive, 1984.

[20] K. T. Poole and H. Rosenthal, "A spatial model for legislative roll call analysis," *American Journal of Political Science*, pp. 357–384, 1985.

[21] J. J. Heckman and J. M. Snyder Jr, "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators," tech. rep., National bureau of economic research, 1996.

[22] S. Jackman, "Multidimensional analysis of roll call data via bayesian simulation: Identification, estimation, inference, and model checking," *Political Analysis*, vol. 9, no. 3, pp. 227–241, 2001.

[23] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," *American Political Science Review*, vol. 98, no. 02, pp. 355–370, 2004.

[24] J. D. Clinton and S. Jackman, "To simulate or NOMINATE?," *Legislative Studies Quarterly*, vol. 34, no. 4, pp. 593–621, 2009.

[25] S. A. Jessee, *Ideology and spatial voting in American elections*. Cambridge University Press, 2012.

[26] M. Kellermann, "Estimating ideal points in the British House of Commons using early day motions," *American Journal of Political Science*, vol. 56, no. 3, pp. 757–771, 2012.

[27] N. Beauchamp, "Text-based scaling of legislatures: A comparison of methods with applications to the US Senate and UK House of Commons." Unpublished manuscript, available at `http://nicholasbeauchamp.com/work/Beauchamp_scaling_2.24.11.pdf`, 2010.

[28] S. Gerrish and D. M. Blei, "How they vote: Issue-adjusted models of legislative behavior," in *Advances in Neural Information Processing Systems*, pp. 2753–2761, 2012.

[29] A. Ceron, "Brave rebels stay home: Assessing the effect of intra-party ideological heterogeneity and party whip on roll-call votes," *Party Politics*, vol. 21, no. 2, pp. 246–258, 2015.

[30] B. E. Lauderdale and A. Herzog, "Measuring political positions from legislative speech," *Political Analysis*, vol. 24, no. 3, pp. 374–394, 2016.

[31] I. S. Kim, J. Londregan, and M. Ratkovic, "Estimating ideal points from votes and text," *Political Analysis*, vol. 26, no. 2, pp. 210–229, 2018.

[32] S. Moser, A. Rodriguez, and C. L. Lofland, "Multiple ideal points: Revealed preferences in different domains," *Political Analysis*, vol. Forthcoming, 2020.

[33] Adler, E. S., and J. Wilkerson, *Congressional Bills Project [Data File and Codebook]*., 2017.

[34] A. Spirling and I. McLean, "UK OC OK? Interpreting optimal classification scores for the UK House of Commons," *Political Analysis*, vol. 15, no. 1, pp. 85–96, 2007.

[35] A. Spirling and K. Quinn, "Identifying intraparty voting blocs in the UK House of Commons," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 447–457, 2010.

[36] C. F. Karpowitz, J. Q. Monson, K. D. Patterson, and J. C. Pope, "Tea time in America? The impact of the Tea Party movement on the 2010 midterm elections," *PS: Political Science and Politics*, vol. 44, no. 2, pp. 303–309, 2011.

[37] K. Arceneaux and S. P. Nicholson, "Who wants to have a tea party? The who, what, and why of the Tea Party movement," *PS: Political Science & Politics*, vol. 45, no. 4, pp. 700–710, 2012.

[38] T. Skocpol and V. Williamson, *The Tea Party and the remaking of Republican conservatism*. Oxford University Press, 2016.

[39] J. Lewis, "Why are Ocasio-Cortez, Omar, Pressley, and Talib estimated to be moderates by NOMINATE?." `https://voteview.com/articles/Ocasio-Cortez_Omar_Pressley_Tlaib`, 2019.

[40] J. Lewis, "Why is Alexandria Ocasio-Cortez estimated to be a moderate by NOMINATE?." `https://voteview.com/articles/ocasio_cortez`, 2019.

[41] R. Guimerà and M. Sales-Pardo, "Justice blocks and predictability of US Supreme Court votes," *PloS one*, vol. 6, no. 11, 2011.

[42] H. Crane, "A hidden Markov model for latent temporal clustering with application to ideological alignment in the US Supreme Court," *Computational Statistics & Data Analysis*, vol. 110, pp. 19–36, 2017.

[43] R. Carroll, J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal, "The structure of utility in spatial models of voting," *American Journal of Political Science*, vol. 57, no. 4, pp. 1008–1028, 2013.

[44] M. Humphreys and M. Laver, "Spatial models, cognitive metrics, and majority rule equilibria," *British Journal of Political Science*, vol. 40, no. 1, pp. 11–30, 2010.

[45] J. X. Eguia, "Challenges to the standard Euclidean spatial model," in *Advances in Political Economy*, pp. 169–180, Springer, 2013.

[46] J. Duck-Mayr and J. M. Montgomery, "Ends against the middle: Scaling votes when ideological opposites behave the same for antithetical reasons," tech. rep., Department of Political Science, Washington University in St. Louis, 2020.

[47] H. F. Weisberg, "Dimensionland: An excursion into spaces," *American Journal of Political Science*, pp. 743–776, 1974.

[48] F. J. Pierre, "Le siècle des idéologies," 2002.

[49] J. Taylor, *Where did the party go?: William Jennings Bryan, Hubert Humphrey, and the Jeffersonian legacy*. University of Missouri Press, 2006.

[50] J. M. Davis, "The transitivity of preferences," *Systems Research and Behavioral Science*, vol. 3, no. 1, pp. 26–33, 1958.

[51] A. Tversky, "Intransitivity of preferences," *Preference, Belief, and Similarity*, p. 433, 1969.

[52] A. Rodríguez and S. Moser, "Measuring and accounting for strategic abstentions in the us senate, 1989–2012," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 64, no. 5, pp. 779–797, 2015.

[53] D. J. Bartholomew, "Scaling binary data using a factor model," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 46, no. 1, pp. 120–123, 1984.

[54] J. M. Legler, L. M. Ryan, and L. M. Ryan, "Latent variable models for teratogenesis using multiple binary outcomes," *Journal of the American Statistical Association*, vol. 92, no. 437, pp. 13–20, 1997.

[55] A. Ansari and K. Jedidi, "Bayesian factor analysis for multilevel binary observations," *Psychometrika*, vol. 65, no. 4, pp. 475–496, 2000.

[56] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[57] J. F. Geweke and K. J. Singleton, "Maximum likelihood" confirmatory" factor analysis of economic time series," *International Economic Review*, pp. 37–54, 1981.

[58] J. S. Liu and Y. N. Wu, "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1264–1274, 1999.

[59] D. B. Rubin and N. Thomas, "Using parameter expansion to improve the performance of the em algorithm for multidimensional irt population-survey models," in *Essays on item response theory*, pp. 193–204, Springer, 2001.

[60] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[61] J. Bafumi, A. Gelman, D. K. Park, and N. Kaplan, "Practical issues in implementing and understanding bayesian ideal point estimation," *Political Analysis*, vol. 13, pp. 171–87, 2005.

[62] M. S. Bartlett, "Comment on 'a statistical paradox' by dv lindley," *Biometrika*, vol. 44, no. 1-2, pp. 533–534, 1957.

[63] T. L. Griffiths and Z. Ghahramani, "The indian buffet process: An introduction and review.," *Journal of Machine Learning Research*, vol. 12, no. 4, 2011.

[64] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, pp. 113–162, 2011.

[65] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, A. Stuart, *et al.*, "Optimal tuning of the hybrid monte carlo algorithm," *Bernoulli*, vol. 19, no. 5A, pp. 1501–1534, 2013.

[66] M. Bédard, "Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234," *Stochastic Processes and Their Applications*, vol. 118, no. 12, pp. 2198–2222, 2008.

[67] M. Betancourt, S. Byrne, and M. Girolami, "Optimizing the integrator step size for hamiltonian monte carlo," *arXiv preprint arXiv:1411.6669*, 2014.

[68] M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 2, pp. 123–214, 2011.

[69] S. Byrne and M. Girolami, "Geodesic monte carlo on embedded manifolds," *Scandinavian Journal of Statistics*, vol. 40, no. 4, pp. 825–845, 2013.

[70] Z. Wang, S. Mohamed, and N. Freitas, "Adaptive hamiltonian and riemann manifold monte carlo," in *International Conference on Machine Learning*, pp. 1462–1470, 2013.

[71] X. Lu, V. Perrone, L. Hasenclever, Y. W. Teh, and S. J. Vollmer, "Relativistic monte carlo," *arXiv preprint arXiv:1609.04388*, 2016.

[72] Y. Zhang, X. Wang, C. Chen, R. Henao, K. Fan, and L. Carin, "Towards unifying hamiltonian monte carlo and slice sampling," in *Advances in Neural Information Processing Systems*, pp. 1741–1749, 2016.

[73] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc*, vol. 37, no. 3, pp. 81–91, 1945.

[74] P. Diaconis, S. Holmes, M. Shahshahani, *et al.*, "Sampling from a manifold," in *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pp. 102–125, Institute of Mathematical Statistics, 2013.

[75] H. Federer, *Geometric measure theory*. Springer, 2014.

[76] R. Abraham, J. E. Marsden, and J. E. Marsden, *Foundations of mechanics*, vol. 36. Benjamin/Cummings Publishing Company Reading, Massachusetts, 1978.

[77] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[78] A. Gelman and D. Rubin, "Inferences from iterative simulation using multiple sequences.," *Statistical Science*, vol. 7, pp. 457–472, 1992.

[79] J. M. Ragusa and A. Gaspar, "Where's the tea party? an examination of the tea party's voting behavior in the house of representatives," *Political Research Quarterly*, vol. 69, no. 2, pp. 361–372, 2016.

[80] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 64, no. 4, pp. 583–639, 2002.

[81] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde, "The deviance information criterion: 12 years on," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 3, pp. 485–493, 2014.

[82] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, vol. 24, no. 6, pp. 997–1016, 2014.

[83] I. L. Dryden, "Statistical analysis on high-dimensional spheres and shape spaces," *Ann. Statist.*, vol. 33, pp. 1643–1665, 08 2005.

[84] K. T. Poole and H. Rosenthal, *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand, 2000.

[85] C. Hare and K. T. Poole, "The polarization of contemporary american politics," *Polity*, vol. 46, no. 3, pp. 411–429, 2014.

[86] C. Bingham and K. Mardia, "A small circle distribution on the sphere," *Biometrika*, vol. 65, no. 2, pp. 379–389, 1978.

[87] B. Kim, S. Huckemann, J. Schulz, and S. Jung, "Small-sphere distributions for directional data with application to medical imaging," *Scandinavian Journal of Statistics*, vol. 46, no. 4, pp. 1047–1071, 2019.

[88] A. Agresti, *Categorical Data Analysis*. John Wiley & Sons, 2013.

[89] A. S. Fullerton and J. Xu, *Ordered regression models: Parallel, Partial, and non-parallel alternatives*. CRC Press, 2016.

[90] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[91] A. Agresti, *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons, 2010.

[92] T. J. McKinley, M. Morters, J. L. Wood, *et al.*, "Bayesian model choice in cumulative link ordinal regression models," *Bayesian Analysis*, vol. 10, no. 1, pp. 1–30, 2015.

[93] W. H. Greene and D. A. Hensher, *Modeling ordered choices: A primer*. Cambridge University Press, 2010.

[94] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *2009 Ninth international conference on intelligent systems design and applications*, pp. 283–287, IEEE, 2009.

[95] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. A. Gutiérrez, "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification," *Neurocomputing*, vol. 135, pp. 21–31, 2014.

[96] M. Khan, S. Mohamed, B. Marlin, and K. Murphy, "A stick-breaking likelihood for categorical data analysis with latent gaussian models," in *Artificial Intelligence and Statistics*, pp. 610–618, 2012.

[97] M. L. Corstjens and D. A. Gautschi, "Formal choice models in marketing," *Marketing Science*, vol. 2, no. 1, pp. 19–56, 1983.

[98] K. V. Mardia and P. E. Jupp, *Directional statistics*. Wiley, 1999.

[99] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: frequentist and bayesian strategies," *Annals of Statistics*, pp. 730–773, 2005.

[100] A. Bhattacharya and D. B. Dunson, "Sparse bayesian infinite factor models," *Biometrika*, pp. 291–306, 2011.

[101] K. G. Jöreskog and I. Moustaki, "Factor analysis of ordinal variables: A comparison of three approaches," *Multivariate Behavioral Research,*, vol. 36, no. 3, pp. 347–387, 2001.

[102] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, vol. 37. CRC Press, 1989.

[103] J. Aitchison and S. D. Silvey, "The generalization of probit analysis to the case of multiple responses," *Biometrika*, vol. 44, no. 1/2, pp. 131–140, 1957.

[104] I. Moustaki, "A latent variable model for ordinal variables," *Applied psychological measurement*, vol. 24, no. 3, pp. 211–223, 2000.

[105] S. Watanabe, "A widely applicable Bayesian information criterion," *Journal of Machine Learning Research*, vol. 14, no. Mar, pp. 867–897, 2013.

[106] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strategic management journal*, vol. 17, no. 6, pp. 441–458, 1996.

[107] R. L. Thorndike, "Who belongs in the family?," *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[108] Stan Development Team, "RStan: the R interface to Stan," 2019. R package version 2.19.1.

[109] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[110] D. Maclaurin, D. Duvenaud, M. Johnson, and R. P. Adams, "Autograd: Reverse-mode differentiation of native Python," 2015.

[111] I. L. Dryden and K. V. Mardia, *Statistical shape analysis: with applications in R*, vol. 995. John Wiley & Sons, 2016.

[112] S. Huckemann, T. Hotz, and A. Munk, "Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions," *Statistica Sinica*, pp. 1–58, 2010.

[113] C. Ley and T. Verdebout, *Modern directional statistics*. CRC Press, 2017.

[114] R. Guhaniyogi and A. Rodriguez, "Joint modeling of longitudinal relational data and exogenous variables," *Bayesian Anal.*, vol. 15, pp. 477–503, 06 2020.

[115] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[116] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 896–911, 2012.

[117] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10, ACM, 2016.

[118] C. C. Holmes, L. Held, *et al.*, "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian analysis*, vol. 1, no. 1, pp. 145–168, 2006.

[119] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using pólya–gamma latent variables," *Journal of the American statistical Association*, vol. 108, no. 504, pp. 1339–1349, 2013.

[120] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[121] S. Lillioja, D. M. Mott, M. Spraul, R. Ferraro, J. E. Foley, E. Ravussin, W. C. Knowler, P. H. Bennett, and C. Bogardus, "Insulin resistance and insulin secretory dysfunction as precursors of non-insulin-dependent diabetes mellitus: prospective studies of pima indians," *New England Journal of Medicine*, vol. 329, no. 27, pp. 1988–1992, 1993.

[122] I. Chavel, *Riemannian geometry: a modern introduction*, vol. 98. Cambridge university press, 2006.

[123] J. Vanlier, C. A. Tiemann, P. A. Hilbers, and N. A. van Riel, "An integrated strategy for prediction uncertainty analysis," *Bioinformatics*, vol. 28, no. 8, pp. 1130–1135, 2012.

[124] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer, "Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability," *Phil. Trans. R. Soc. A*, vol. 371, no. 1984, p. 20110544, 2013.

[125] M. D. Hoffman and A. Gelman, "The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.

[126] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[127] S. Watanabe, "Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3571–3594, 2010.

[128] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[129] A. Armagan, D. B. Dunson, and J. Lee, "Generalized double pareto shrinkage," *Statistica Sinica*, vol. 23, no. 1, p. 119, 2013.

[130] R. B. O'Hara, M. J. Sillanpää, *et al.*, "A review of bayesian variable selection methods: what, how and which," *Bayesian analysis*, vol. 4, no. 1, pp. 85–117, 2009.

[131] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, pp. 711–732, 1995.

[132] T. Chen, E. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *International Conference on Machine Learning*, pp. 1683–1691, 2014.

[133] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven, "Bayesian sampling using stochastic gradient thermostats," in *Advances in neural information processing systems*, pp. 3203–3211, 2014.

[134] A. Agresti, "Analysis of ordinal categorical data," 1984.

[135] J. Johndrow, D. Dunson, and K. Lum, "Diagonal orthant multinomial probit models," in *Artificial Intelligence and Statistics*, pp. 29–38, 2013.

[136] E. M. Penn, "A model of farsighted voting," *American Journal of Political Science*, vol. 53, no. 1, pp. 36–54, 2009.

[137] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney, "Spherical topic models," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 903–910, 2010.

[138] A. Criminisi, J. Shotton, E. Konukoglu, *et al.*, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, 2012.

[139] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[140] J. Wang, Z. Zhang, and H. Zha, "Adaptive manifold learning," in *Advances in neural information processing systems*, pp. 1473–1480, 2005.

[141] M. Laver, "Measuring policy positions in political space," *Annual Review of Political Science*, vol. 17, pp. 207–223, 2014.

[142] J. Jang and D. B. Hitchcock, "Model-based cluster analysis of democracies.," *Journal of Data Science*, vol. 10, 2012.

[143] A. Ceron, "The politics of fission: An analysis of faction breakaways among italian parties (1946–2011)," *British Journal of Political Science*, vol. 45, no. 1, pp. 121–139, 2015.

[144] J.-F. Godbout and B. Høyland, "Legislative voting in the canadian parliament," *Canadian Journal of Political Science/Revue canadienne de science politique*, vol. 44, no. 2, pp. 367–388, 2011.

[145] A. Ceron, "Brave rebels stay home: Assessing the effect of intra-party ideological heterogeneity and party whip on roll-call votes," *Party Politics*, vol. 21, no. 2, pp. 246–258, 2015.

[146] M. Betancourt, "A conceptual introduction to hamiltonian monte carlo," *arXiv preprint arXiv:1701.02434*, 2017.

[147] M. Betancourt and M. Girolami, "Hamiltonian monte carlo for hierarchical models," *Current trends in Bayesian methodology with applications*, vol. 79, p. 30, 2015.

[148] K. T. Poole and H. Rosenthal, "The polarization of the congressional parties," *Voteview. com, Updated March*, vol. 21, p. 2015, 2015.

[149] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 51–58, 2019.

[150] S. Hauberg, "Principal curves on riemannian manifolds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1915–1921, 2015.

[151] S. Huckemann and H. Ziezold, "Principal component analysis for riemannian manifolds, with an application to triangular shape spaces," *Advances in Applied Probability*, vol. 38, no. 2, p. 299âĂŞ319, 2006.

[152] Policy Agendas Project, *Roll-Call Votes*, 2017.

[153] J. C. Chan and A. L. Grant, "Fast computation of the deviance information criterion for latent variable models," *Computational Statistics & Data Analysis*, vol. 100, pp. 847–859, 2016.

[154] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *Journal of the american Statistical association*, vol. 100, no. 469, pp. 286–295, 2005.

[155] A. D. Martin and K. M. Quinn, "Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999," *Political analysis*, vol. 10, no. 2, pp. 134–153, 2002.