Modeling Simplex-valued Data and Latent Structures

by

Rayleigh Lei

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Statistics) in the University of Michigan 2022

Doctoral Committee:
Professor Long Nguyen, Chair
Assistant Professor Yang Chen
Professor Fred Feinberg
Assistant Professor Jeffrey Regier
DEDICATION

To my family and girlfriend, because with their love and support, all is possible.
ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation to my advisor, Dr. Long Nguyen, for his insightful guidance, sustained inspiration, and strong support during my PhD study. His mentorship has helped nourish and shape my growth and development as an effective researcher and junior scholar in the statistical field. I am also most grateful to all my committee members, Dr. Fred Feinberg, Dr. Jeff Regier, and Dr. Yang Chen, for always keeping an open door and providing invaluable advice and feedback to my research and generous support to my career.

Many outstanding collaborators have tremendously aided the completion of my PhD research. Dr. Elizabeth Bruch introduced me to the Los Angeles income proportions data set. The data set, her professional inputs, and Lydia Wileden’s help in preparing the data were a vital part of my thesis. I would like to thank Dr. Áine Heneghan, Dr. Nathan Martin, Dr. Rene Rusch, and Dr. Sam Mukherji for the opportunity to explore music theory with statistics. This inspired me to investigate a model that is the backbone of one of the chapters in my thesis. Their constructive conversations enabled me to learn and understand many new concepts related to the structure of classical music. My sincere appreciation goes to Dr. Ding Zhao and his lab for providing interesting data sets and demonstrating the power of optimal transport and Bayesian nonparametrics to better analyze these data sets. Our meetings always featured fascinating new ideas to comprehend and new challenges to overcome.

I would like to thank my lab mates and the undergraduate students who I mentored and taught. Exchanging ideas with them in various settings and collaborating with them have been a boon to my research and enriched my learning and graduate school life.

Finally, I would like to thank my entire family, friends, and girlfriend for their unflagging encouragement, unconditional support, and endless love over the years.

As I finish up my Ph.D. study at University of Michigan, I fully recognize that I could not have accomplished this by myself. It is only through a community, such as the amazing one described above, that I have been able to. I am truly blessed to have been associated with this community during my Ph.D. study and hope this continues in my future career and life.

My thesis research has been supported by the NSF Graduate Research Fellowship Program (Grant No. DGE 1256260). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.
# TABLE OF CONTENTS

DEDICATION ........................................................................................................... ii

ACKNOWLEDGMENTS ........................................................................................... iii

LIST OF FIGURES ................................................................................................. vii

LIST OF TABLES .................................................................................................. xi

LIST OF APPENDICES ......................................................................................... xiii

LIST OF ACRONYMS ............................................................................................ xiv

ABSTRACT ............................................................................................................. xv

CHAPTER

1 Introduction ......................................................................................................... 1

2 Simplex Movement ............................................................................................ 8

2.1 Introduction ................................................................................................... 8

2.2 Affine transformation .................................................................................... 9

2.3 Error Models ................................................................................................ 11

2.3.1 Additive Dirichlet error model ......................................................... 12

2.3.2 Exponential rotational error model ............................................... 15

2.3.3 Rotational geodesic error model ..................................................... 16

2.4 Simulation studies ....................................................................................... 20

2.4.1 Additive Dirichlet error model ....................................................... 20

2.4.2 Exponential rotational error model ............................................... 21

2.4.3 Rotational geodesic error model ..................................................... 24

2.5 Data analysis ................................................................................................ 29

2.6 Conclusion .................................................................................................... 32

3 Simplex Random Direction ................................................................................ 33

3.1 Introduction ................................................................................................... 33

3.2 Polar coordinate transformations for circular data ...................................... 36

3.2.1 von Mises distributions .................................................................... 37

3.2.2 Projected normal distribution .......................................................... 39
3.2.3 Projected Gaussian process ........................................... 40
3.2.4 Higher dimension analogues ........................................ 41
3.3 Modeling random directions .......................................... 42
  3.3.1 Model description .................................................. 42
  3.3.2 Model properties for spatially dependent von Mises models ... 47
3.4 Posterior inference ..................................................... 51
3.5 Simulations and results ................................................. 54
  3.5.1 Simulation study ................................................... 54
  3.5.2 Data analysis ...................................................... 58
3.6 Conclusion and future directions ................................. 63

4 Tree-based Mixture of Probability Simplices .......................... 65
4.1 Introduction ............................................................ 65
4.2 Background ........................................................... 68
4.3 Model ................................................................. 70
  4.3.1 Notation and model setting ...................................... 70
  4.3.2 Metrics on the tree-based mixtures ............................. 72
  4.3.3 Identifiability and consistency ................................ 74
4.4 Model fitting and computation ...................................... 81
  4.4.1 Model ............................................................ 81
  4.4.2 Gibbs update ..................................................... 82
  4.4.3 Estimating using Variational Bayes ............................ 83
  4.4.4 Geometric algorithms .......................................... 86
  4.4.5 Metric computation ............................................ 90
4.5 Simulation and data analysis ....................................... 92
  4.5.1 Simulation ........................................................ 92
  4.5.2 New York Times corpus ......................................... 95
4.6 Conclusion ............................................................ 99

5 Conclusion ............................................................... 101

APPENDICES .............................................................. 105

A Simplex Movement Appendix ........................................... 105
  A.1 Affine transformation properties .................................. 105
  A.2 Additive Dirichlet error model .................................... 106
    A.2.1 $\gamma$ log likelihood gradient .............................. 107
    A.2.2 $\eta$ log likelihood gradient ................................ 108
    A.2.3 Multiple direction affine transformation with additive Dirichlet error ..................................... 110
  A.3 Rotational geodesic error model ................................... 120
    A.3.1 Geodesics in $\Delta^2$ .......................................... 120
    A.3.2 Generating the model in $\Delta^2$ ............................ 122
    A.3.3 Geodesics in higher dimensions .............................. 126

B Simplex Random Direction Appendix ................................ 134
LIST OF FIGURES

FIGURE

2.1 Left plot shows the affine transformation when $\eta$ is negative whereas the right plot shows the transformation when $\eta$ is positive. ............................................ 9

2.2 Plots showing affine transformed data on the left and affine transformed data with the additive Dirichlet error model on the right ($\gamma = 0.1, \alpha = 2$). Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$. ................................................................. 20

2.3 Optimization results for different starting values of $\eta$ and $\gamma$ plotted on a heatmap of the log likelihood likelihood. The grey areas are areas in which the log likelihood is not defined. The x-axis represents different values of $\eta$ and the y-axis represents different values of $\gamma$. ......................................................... 21

2.4 Plots showing affine transformed data on the left and affine transformed data with the exponential rotational error model, $\lambda = 10$, on the right. Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$. ......................................................... 22

2.5 Plots of the log likelihood for $\eta = -0.2$ and $\eta = 0.1$ and the expected end points given a direction for different values of $\lambda$. The x-axis represents different values of $\eta$. ......................................................... 23

2.6 Plots showing the posterior mean and credible intervals for $\eta$. The left two plots show simulation results under different model when $\eta$ is known and shown by the yellow line. The right plot are the estimates for the data transformation from one year to the next connected for visualization purposes. .......................................................... 24

2.7 Plots showing affine transformed data on the left and affine transformed data with the rotational geodesic error model with $\psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38$ and Beta(1, 3) for the proportion probability on the right. Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$. ................................................................. 25

2.8 Plots of the log likelihood for $\eta = -0.2$ and $\eta = 0.1$ under the rotational geodesic error model with $\psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38$ and Beta(1, 3) for the proportion probability. The x-axis represents different values of $\eta$. ......................................................... 25

2.9 Left plots show the likelihood contour for $v_1, v_2$, and $\eta \in [-0.5, 0.5]$ with the log likelihood at $(1, 0, 0)$ removed whereas the right plots show the likelihood surface for $\eta \in [-0.5, 0.5]$ along the line $v = (0, 1/2, 1/2)(1 - t) + (1, 0, 0)t$. The value used to simulate the data are $v = (1, 0, 0)$ and $\eta = -0.2$ for the top plots and $\eta = 0.1$ for the bottom plots. $\psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38$ and Beta(1, 3) for the proportion probability were used to calculate the log probability. ................................................................. 26

2.10 Trace plots of $\eta$ for $v_i = 1$ if $i$ is the true direction and $v_i = 0$ if $i$ is not the direction. The left plot shows results that start from $(1/3, 1/3, 1/3)$ and the right plot shows results from $(0.8, 0.1, 0.1)$. Yellow line indicates the true value, $\eta = 0.1$. ................................................................. 27
2.11 Plots of the log likelihood under the affine transformation with \( \eta = 0.1 \) and the rotational geodesic error model with Beta(1, 3) for the proportion probability and different values of \( \psi_0, \psi_1, \psi_2 \). The x-axis represents different values of \( \eta \).

2.12 Plots of the scaled posterior for \( \eta = -0.2 \) and \( \eta = 0.1 \) and \( v = (1/3, 1/3, 1/3) \) after running MCMC for different values of \( b \in [2.5, 3.5] \). \( b \) is the parameter from Beta(1, \( b \)), which assigns probability to the proportions.

2.13 Plots showing the results of running the MCMC sampler and sampling all indices of \( v \) before sampling \( \eta \), starting at \( (1/3, 1/3, 1/3) \) unless otherwise indicated.

2.14 Plots showing the results of running the MCMC sampler and sampling an index of \( v \) and then sampling \( \eta \), starting at \( (1/3, 1/3, 1/3) \) unless otherwise indicated.

3.1 Random direction plots for different years. The \( x-y \) coordinates represent the income proportion in the first two categories. If the plot is three dimensional, the \( z \) coordinate is the random direction divided by \( 2\pi \) to lie in \([0, 1]\). The two leftmost plots show all random directions for the years listed below. The middle plots show the movement for a subset of 20-30 “locations” and the corresponding extracted directions. The two rightmost plots display a histogram of the random directions associated with locations that are within \( 0.05 \ L_2 \)-distance of \((0.39, 0.56, 0.05)\) for 2003-2004 and of \((0.85, 0.10, 0.05)\) for 2008-2009.

3.2 Plots showing the elliptical slice sampler’s proposal ellipse for different locations (shown as circles) and means for the Gaussian prior (shown as the squares), but the same elliptical proposal point (shown as the triangle). These ellipses represent the set of angles that can be proposed by the elliptical slice sampler. The arrow demonstrates that the locations have the same angle in polar coordinates, but different radii.

3.3 The right three plots show the mean component fitted by different models for random directions simulated according to \( SvM-c \) in (3.15). The fitted mean components for the different models are shown in colored asterisks with the simulated random directions as circles and the simulated means as grey rhombuses. The leftmost plot is just the simulated data and the mean surface used to generate that data. The \( x-y \) axis show the locations in the first two dimensions of a three dimensional simplex; the \( z \)-axis represent the direction.

3.4 Plots showing the change in income proportion for census tract 6037920336 and its neighboring tract on the left and the random direction of the movements for 2003-2004 by census tract on the right. The center of the census tract in longitude and latitude is used to represent the census tract.

3.5 Plots showing the mean direction and credible intervals averaged over all locations for various hyperparameters for \( SvM-c \) compared to \( \omega = 0.1 \) and \( \sigma = 0.5 \) on the left and showing the probability and credible intervals averaged over all locations for various hyperparameters for \( SvM-p \) compared to \( \omega = 0.1 \) and \( \sigma = 1 \) on the right. The value for \( \omega \) is given first.
3.6 Plots showing the observed random direction not withheld and the fitted mean surface of the model selected by the posterior predictive log probability in (3.34). The front axis represents the proportion in the first income category and the side axis represents the proportion in the second income category. This is reversed in the alternative view. Regardless, the up-down axis represents the direction. The start of the arrow indicates a value of zero whereas the end indicates a value of $2\pi$ for angles and 1 for proportions.

4.1 Example of a tree-based mixture of probability simplex. A path, $\varphi_d$, is shown in blue. 70

4.2 Plots showing the multidimensional scaling representation of the simplices estimated by various methods. Because the labels are invariant under a permutation, the simplices are assigned so that the total Hausdorff metric between the simulated simplices and estimated simplices are minimized. 94

4.3 Plots showing the projected simplicial distance defined in (4.4.1) versus the log likelihood used in the AIC and BIC calculation for different vocabulary sizes. The plots were generated from counts from a Multinomial distribution with $n = 200$ and an uniform probability. The mean proportions were projected onto a smaller simplex with vertices, $\theta^i$, such that $\theta^i = \frac{1}{D-1}$ and $\theta^j = \frac{D-2}{(D-1)^2}$ for $i, j \in 1, 2, \ldots, D, i \neq j$. 96

4.4 Best fitted topic tree showing the five most likely words associated with each topic from Variational Bayes according to the criterion listed below. The number of documents associated with each path is also displayed at the bottom of the tree. 98

A.1 Plots comparing the simplex Hellinger Distance proportion against $t \in [0, 1]$ for different $\theta_1$. 126

B.1 Plots in the top row show the duplicate observations broken down by different categories across different years with further context provided by plots in the bottom row. 174

B.2 Plots showing the posterior predictive log probability defined in (3.34) and the von Mises distributions’ parameters and the mixing probability of the first component selected according to that probability and the models’ mixing for each year. For $SvM$ and $SvM-c$, the mean and concentration parameters’ mean and credible interval are averaged over all locations. On the other hand, the mean and credible interval for the probability of the first component is averaged across all location for $SvM-p$. In addition, the second component’s values are shown with an opacity of 0.5. 174

B.3 Plots showing the observed random direction not withheld and the fitted mean surface of the model selected by the posterior predictive log probability in (3.34). The front axis represents the proportion in the first income category and the side axis represents the proportion in the second income category. The up-down axis represents the direction. The start of the arrow indicates a value of zero whereas the end indicates a value of $2\pi$ for angles and 1 for proportions. 175

B.4 Plots of the mean component fitted by different models for observations simulated according to the $SvM$ model in (B.3). The fitted mean components are shown in colored asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations. 177
B.5 Plots of the mean component fitted by different models for observations simulated according to the $SvM$ model in (B.3). The fitted mean components are shown in colored asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations.  

B.6 Plots of the mean component fitted by different models for observations simulated according to the $SvM-p$ model (B.3) in the bottom row. The fitted mean components are shown in red asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations.  

C.1 Best fitted topic tree of the five runs showing the five most likely words associated with each topic from various methods according to the projected simplicial distance. The number of documents associated with each path is also displayed at the bottom of the tree. The left one is derived from geometric method whereas the right one is from VB.
LIST OF TABLES

TABLE

2.1 Mean posterior estimates for $\eta$ and $v$, starting at $(1/3, 1/3, 1/3)$. Results from sampling all indices of $v$ before $\eta$ are on the left and Results from sampling an index of $v$ and then $\eta$ are on the right. ......................................................... 32

3.1 A table showing the relationship between the models discussed in Section 3.3.1 and the supplementary material. The models introduced in this chapter are in bold. . . . . 43

3.2 Circular mean posterior values and 95% credible intervals for the parameters in von Mises distribution and mixing probability are shown for the simulation data in Figure 3.3. Parameters with bars over them were averaged across all locations. . . . . . . 56

3.3 The log of the posterior predictive probability given in Equation (3.34) for 50 draws averaged across posterior draws and then posterior predictive draws. . . . . . . . . 57

4.1 Various distances between vertices used to generate the simulation and vertices as estimated by different methods for data of distinct dimension and in parentheses, the average projected simplicial distance with the minimum distance across all methods in bold. The first line presents the minimized sum of $L_2$ distance between topics. The next two lines show the path and tree metric respectively defined in (4.3.2) and (4.3.4) and computed in the manner discussed in Section 4.4.5. The last line illustrates the projected simplicial distance (4.4.1). The table on the left shows results when the tree size is fixed so that the tree is three deep and each node except the leaf nodes has two children and $D$ varies whereas the table on the right displays the results when $D$ is fixed to be 100 and the tree size varies. . . . . . . . . . . . . . . . . . . . . . . . . . 93

B.1 Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the income proportions in Los Angeles County from 1990 to 2010. Parameters with bars over them were averaged across all locations. . . . . . . . . . . . . . . . . . . . . . . . . . 173

B.2 Time it takes for our models to run four chains for different models in seconds according to the sampling scheme outlined in Section 3.5.1 on four cores of a cluster using 2x 3.0 GHz Intel Xeon Gold 6154 as its processor. . . . . . . . . . . . . . . . . . . 176

B.3 Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.4. Parameters with bars over them were averaged across all locations. . . . . . . . . . . . . . . . . . . . . . . . . . 177
B.4 Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.4. Parameters with bars over them were averaged across all locations. ................................................................. 178

B.5 Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.6. Parameters with bars over them were averaged across all locations. ................................................................. 179

C.1 Time in seconds that it took to run the various approaches for data of different dimensions. For the Gibbs sampler, 4 chains with 2000 iterations were run in parallel. .... 194

C.2 Time in seconds that it took to run the various approaches for different vocabulary sizes. For the Gibbs sampler, 4 chains with 2000 iterations were run in parallel. .... 194

C.3 Time in seconds that it took to run the various approaches for different vocabulary sizes. For the geometric approaches, the times were averaged across five runs. .... 195
LIST OF APPENDICES

A Simplex Movement Appendix ............................................. 105
B Simplex Random Direction Appendix ........................................ 134
C Tree-Based Mixture of Simplices Appendix ................................. 180
LIST OF ACRONYMS

alr  additive log-ratio
clr  centered log-ratio
HMDA  Housing Mortgage Disclosure Act
i1r  isometric log-ratio
LDA  Latent Dirichlet Allocation
SvM  spatially varying von Mises model
SvM-c spatially varying von Mises component mixture model
SvM-p  spatially varying mixing probability for von Mises distributions model
VB  Variational Bayes
Tree VB  Tree Variational Bayes
Examples of data that lie in a simplex abound in a variety of fields. The chemical, mineral, and/or fossil percentages of a collection of rocks is of interest to geologists. Demographers and policy makers might examine the income proportions or racial compositions of neighborhoods or other political units. However, it can be challenging to model this type of data, particularly because each observation must sum up to one. If not handled correctly, this might induce what Karl Pearson called "spurious correlation". As a trivial example of this, take an observation that has two values and sums up to one. Then, the values must be negatively correlated with each other. While traditional approaches for these type of data involve analyzing the log ratio transforms, this might be problematic if any of the observations have a zero as one of their values or for interpretation.

Additional challenges arise if this data changes over time. The first two chapters lay the framework for how such data may be modeled. The second chapter proposes doing so using a general affine transformation for the overall change and a sufficiently rich error model for the difference between the overall transformation and the observation at the next time point. Of the three models explored, the rotational geodesic error model is most promising. However, it might not be appropriate to assume that the direction observations moved in is uniformly distributed. Using ideas from directional statistics, we discuss in the third chapter how to model directions that appear to be similar for observations with similar values. In both chapters, we run simulation studies and analyze the income proportions from Los Angeles County. In each case, our analysis is able to discover trends consistent with larger macroeconomic ones and provide further details about these trends.

The last chapter discusses tree-based mixtures of probability simplices. In other words, the simplices share vertices in a way that can be represented by a tree with the root node corresponding to a vertex shared by all simplices and the leaf nodes corresponding to vertices present in one simplex. We show when such models have posterior consistency and demonstrate how to efficiently fit them using geometric methods. Indeed, we apply them to analyze a subset of articles from the New York Times, uncovering meaningful topics and interesting semantic relationships between these topics. While we leave it to future work, these methods might also be combined with the ones from previous chapters to model how sub-regions of data that lie in a simplex change over time.
CHAPTER 1

Introduction

Some or all the primary quantities of interest in many real-world problems are non-negative proportions that sum up to one and thus lie in a probability simplex. For instance, to study a collection of rocks and sediments, geologists might analyze the chemical, mineral, and/or fossil percentages of each sample to determine the effects of various natural processes or which samples are related [2]. As another example, microbiologists might examine microbiome data from a set of volunteers because of the microbiome’s effect on human health [7]. Here, the microbiome refers to the proportions of various microbes in the human gut [7]. Social sciences also provide examples. Demographers or policymakers might be interested in understanding how the income proportions or the racial composition of neighborhoods, census tracts, cities, or other political units change over time in order to better understand the effects of policies or economic events. Alternatively, researchers might examine the proportion of words from documents across a corpus to find and understand topics, i.e., distributions of words [74, 75, 76].

Modeling data that lie on the simplex can be challenging. To be concrete, we let \( \Delta^D \) denote the \( D \)-dimensional probability simplex, i.e., the subset of elements in \( \mathbb{R}^{D+1} \) whose components are non-negative and sum to one. Given a data set represented by a collection of random samples \( x_1, \ldots, x_N \in \Delta^D \), each data point \( x_i \) is composed of the \( D \) components \( x_{i,j} \) for \( j = 1, \ldots, D \). The most immediate difficulty is the simplex constraint imposed by the proportions because they must sum up to one [57]. For this constraint to be maintained, a change in one proportion necessarily involves a change in the other proportions. If this is not handled correctly, the model might detect what Pearson called "spurious correlation" [58]. As a trivial example, take \( \Delta^1 \). The correlation between the two coordinates must be -1 [56].

However, despite interest in this field and this warning from Pearson in 1896, there was no further progress in methodology until the 1980s [5]. Indeed, researchers continued to use standard multivariate analysis or modified versions of these types of analysis despite further criticism in multiple fields around the 1960s concerning how there is a negative bias in the correlation of proportional data due to the proportions needing to sum up to one [14, 48, 68]. Obviously, this could affect standard exploratory data methods that analyze the covariance structure, such as principal
component analysis or factor analysis [56]. However, as discussed by Rock, other statistical techniques and properties might be affected as well [61]. All in all, while reasonable, interpretable results might have been found, there still might be a lingering question whether the results were statistically significant or just spurious correlations [56].

Then, in a series of papers in the 1980s, John Aitchison singlehandedly proposed a new framework to tackle these type of problems [1, 2, 3, 4, 6]. Recognizing that proportions contain relative information about how each component relates to each other component, Aitchison suggested analyzing the log ratios of components to some quantity. The log ratios were chosen instead of the ratios because the log ratios are easier to work with and define a one-to-one mapping to the real space so long as no components are zero. In particular, Aitchison introduced two such transformations [2, 4]. First, the additive log-ratio (alr) is the following transformation for \( \mathbf{x}_i \) from \( \Delta^D \) to \( \mathbb{R}^D \):

\[
(alr(\mathbf{x}_i))_j = \log\left( \frac{x_{i,j}}{x_{i,D}} \right) \quad j = 1, 2, \ldots, D - 1.
\]

In other words, it takes the log transform of the ratio of any component except the last to the last. Alternatively, given \( x \in \Delta^D \), the centered log-ratio (clr) is the following transformation from \( \Delta^D \) to \( \mathbb{R}^{D+1} \):

\[
(clr(\mathbf{x}_i))_j = \log\left( \frac{x_{i,j}}{g(\mathbf{x}_i)} \right) \quad j = 1, 2, \ldots, D.
\]

Here, \( g(\cdot) \) indicates the geometric mean. As a result, this transformation is a "centered" log transform of the proportions because for \( j = 1, 2, \ldots, D \), \( clr(x_{i,d}) = \log(x_{i,d}) - \frac{1}{D} \sum \log(x_{i,d}) \). Then, once a transformation is chosen, standard multivariate analysis can be applied to these transformed proportions.

While these initial log ratio techniques seemed to be accepted by the statistics community, there were criticisms from other fields [5]. Indeed, there are some limitations with the transformations discussed above [23]. First, with the alr transformation, it is possible that depending on which proportion is used, there might be a different alr transformation. Next, there are issues with interpretation. The ratio we are analyzing compares one component against the last component, but not against any other particular component. Meanwhile, the last component is also being contrasted against all other components. Hence, there is some difficulty in interpreting a particular transformed proportion and understanding what it means [23]. Next, with the clr transformation, notice that the sum must necessarily be zero. Consequently, the covariance matrix is degenerate and the transformed data is still constrained [23]. In addition, the clr has the opposite problem in interpretation as the alr. Due to the geometric mean, each proportion is equally compared against each other proportion, including itself. Further, the geometric mean leads to overlap between

2
the transformed proportions. As a result, the transformed proportions cannot be simultaneously
terpreted as a function of the original proportions [23].

There is also one more limitation that spurred the next significant development in log ratio
transformations. This issue can only be understood in the context of the Aitchison geometry, which
is endowed with an Aitchison inner product, norm, and distance. Here, the Aitchison inner product
is defined to be the usual inner product, but between the log ratio [56]. It is denoted as \( \langle \cdot, \cdot \rangle_a \). In
other words, for \( i, i' \in 1, 2, \ldots, N \),
\[
\langle x_i, x_{i'} \rangle_a = \frac{1}{2D} \sum_{j=1}^{D+1} \sum_{k=1}^{D+1} \log \left( \frac{x_{i,j}}{x_{i,k}} \right) \log \left( \frac{x_{i',j}}{x_{i',k}} \right).
\]
(1.3)

Meanwhile, the norm, \( \| \cdot \|_a \), is simply \( \langle x_i, x_i \rangle_a \). Defining the distance requires introducing two
operations: perturbation and powering. The perturbation, \( \oplus \), of \( x_i \) and \( x_{i'} \) is defined as the following
operation from \( \Delta^D \times \Delta^D \) to \( \Delta^D \):
\[
(x_i \oplus x_{i'})_j = \frac{x_{i,j}x_{i',j}}{\sum_k x_{i,k}x_{i',k}}.
\]
(1.4)

Powering, which is denoted as \( \odot \), is a function of \( x_i \) and \( t \in \mathbb{R} \). Then, the powering of \( x, t \) is
defined as the following operation from \( \Delta^D \times \mathbb{R} \) to \( \Delta^D \):
\[
(x \odot t)_i = \frac{x_{i,t}}{\sum_{i'} x_{i',t}}.
\]
(1.5)

The inverse of the perturbation, \( \ominus \), can be defined using these operations. If \( x_i, x_{i'} \in \Delta^D \), \( x_i \ominus x_{i'} = x_i \ominus (x_{i'} \odot (-1)) \). Then, the Aitchison distance between \( x_i, x_{i'} \in \Delta^D \) is \( d^2_a(x_i, x_{i'}) = \| x_i \ominus x_{i'} \|^2_a \).

Both transformations discussed thus far can be examined using this geometry. While \( \text{alr}(x \oplus y) = \text{alr}(x) + \text{alr}(y) \) and \( \text{alr}(x \odot t) = t \cdot \text{alr}(x) \) for any \( x, y \in \Delta^D, t \in \mathbb{R} \), this is not the case for the
Aitchison inner product, norm, or distance [23]. In other words, \( \langle x, y \rangle_a \neq \langle \text{alr}(x), \text{alr}(y) \rangle \) and
\( \| x \|_a \neq \| \text{alr}(x) \| \) for \( x, y \in \Delta^D \). Here, \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| \) denote the usual inner product and Euclidean
norm respectively. In essence, the alr transformation is not an isometric transformation from the
simplex with the Aitchison measure to the real space with the Euclidean measure [23]. Instead, the
axes for each transformed proportion are at a 60 degree angle [56]. Meanwhile, while it is possible
to show that the clr is an isometric transformation to some real subspace, it is not straightforward
to derive the orthogonal basis for this subspace [21]. If the data is to be transformed, it might be
desirable to have a mapping that can preserve angles and distances from one space to another
because intuition from one space might carry over [21].

To address this limitation, Egozcue et al. proposed the isometric log-ratio (ilr) [21]. Say vectors
\( e_1, e_2, \ldots, e_D \) are Aitchison-orthonormal if for \( d, d' = 1, 2, \ldots, D, D \in \mathbb{N}, \langle e_d, e_{d'} \rangle_a = 0 \) for
$d = d'$ and $\langle e_d, e_{d'} \rangle_a = 1$ for $d \neq d'$. Then, an ilr transformation of $x_i$ is the following transform from $\Delta^D$ to $\mathbb{R}^D$ given Aitchison-orthonormal vectors, $e_1, e_2, \ldots, e_D$:

$$\text{ilr}(x_i) = (\langle x_i, e_1 \rangle_a, \langle x, e_2 \rangle_a, \ldots, \langle x, e_D \rangle_a).$$

(1.6)

Because the space is equipped with an inner product and norm, Gram-Schmidt might be used to find an orthogonal basis. Alternatively, in their paper on ilr transformations, Egozcue et al. explicitly construct one in Proposition 3 based on the eigenvalues of the clr mapping [21]. There has also been subsequent work to find other basis, allowing proportions to be transformed into pivot coordinates [24], symmetric pivot coordinates [39], or balances [20].

Because it can be shown the ilr transformation is isometric, certain statistical methods become straightforward to apply. For instance, if one wants to cluster the data based on the Aitchison distance, one can apply the ilr transform and cluster the transformed data using existing techniques [23]. As another example, suppose that one wants to perform linear regression to predict a proportion [23]. If $t \in \mathbb{R}^K$ and $y, b_0, b_1, \ldots, b_K, \epsilon \in \Delta^D$ for some $K \in \mathbb{N}$, then the multiple regression framework is the following:

$$y = b_0 \oplus \bigoplus_{k=1}^{K} b_k \odot t_k \oplus \epsilon.$$

(1.7)

In other words, the regression coefficients are a set of proportions that are powered by an observation. Then, these altered regression coefficients are perturbed with each other, an intercept proportion, and a noise proportion. Due to the properties of the ilr transformation, this is equivalent to solving the multiple regression problem for the ilr transformed $y, b_0, b_1, \ldots, b_K, \epsilon$. However, despite this, there still might be issues with interpreting the results depending on the basis used [56].

Still, these transformations were also used in Bayesian inference. For instance, Iyengar and Dey fit a multivariate regression model to the Box-Cox transformed proportions [35]. This Box-Cox transformation is a generalization of the alr transform because given $\lambda_d$, $y_d = \text{alr}(x_d)^{\lambda_d - 1}/\lambda_d$ if $\lambda_d \neq 0$ and $\text{alr}(x_d)$ otherwise. As a result, they sample for the intercept, covariates, the correlation matrix, and $\lambda$. As another example, researchers sought to include spatial information when trying to model alr transformed proportions. To model the proportion of newborns based on the mothers’ age in Brazil, Martinez et al. used Bayesian analysis to fit a multivariate regression with a spatial varying intercept term [47]. The spatial term used a conditional autoregressive to borrow strength from nearby microregions of Brazil. Alternatively, Palmer and Douglas sought to spatially correlate the alr transformed proportions of sediments found in water samples in order to make predictions for a spatial grid [55].

Note that up until this point, this introduction has been discussing the domain of "compositional
data" within statistics. However, developments in other fields that have proportional data have revealed additional challenges to this framework. For instance, because the number of species determines the dimension of the simplex in microbiome data, the dimension of the simplex might be large due to the diversity of species. However, there might not be many samples and there is a lot of sparsity in the data set [67]. The latter is particularly troublesome because the log ratio becomes undefined if any covariates are zero. While there are a variety of strategies to deal with zeros [42], including Bayesian ones, [46], some might not be totally appropriate due to there potentially being "structure" that affects sparsity [44]. Certain species might compete against each other or require certain other species to survive. Then, even among the non-zero proportions, there might be a few dominant species. As a result, a lot of the non-zero terms might be close to zero as well [67]. This is further complicated by potential undersampling when sampling the microbiome data. Some of these same issues also arise when discussing topic modeling. There might be a lot of words, but only some of these words might appear in a document. Further, there might be semantic structure that can be exploited to better find associations between words and documents.

There is one additional challenge that has also not been discussed so far. Suppose that there is a collection of proportions observed at various time points. A natural question is to model how such data can change over time. In the framework of log ratio transformation discussed above, one might use an isometric transformation and then apply a time series analysis on the transformed proportions [60]. However, even in the low dimensional case, this still might be problematic if any of the proportions are zero at a time point. Further, it is desirable to have the change be interpretable. To that end, this thesis lays the groundwork for such a model.

In particular, the next two chapters discuss modeling how the income proportions change in the Housing Mortgage Disclosure Act (HMDA) data from 1990 to 2010 in Los Angeles county. HMDA was enacted in 1975 to address growing public concerns over the practice of "redlining" in the 1970s [25]. The term refers to the practice of denying credit to individuals in certain neighborhoods due to demographic factors or the age of the housing stock instead of their creditworthiness. Following a tradition that transparency and public disclosure could serve as tools to address a private practice contrary to the public good, HMDA sought to help address these concerns by requiring financial institutions to reveal certain information on their loans. Over the years, the amount of information revealed has increased to include demographic information.

As a result of this increase, while HMDA data is still relied upon by regulators to ensure that financial institutions are complying with the Community Reinvestment Act and other anti-discriminatory laws, HMDA data’s usefulness extends beyond regulatory purposes [25]. For instance, a quick search reveals papers using HMDA data to explore the impact of race in the mortgage market and the recent housing mortgage foreclosure crisis [16, 62]. As stated in the previous paragraph, the goal of the first two chapters is to model how the income proportions
change in the HMDA data from 1990 to 2010 in Los Angeles county. While there are sixteen income categories in the original data set, the data set dealt with in this thesis will only have three: $0-$100,000, $100,000-200,000, and $200,000+.

Then, in chapter 2, an affine transformation is combined with a sufficiently rich error model to model how the set of proportions changes from one year to the next. We can track how the proportions change from one year to the next using the census tract IDs. As we will show in chapter 2, the affine transformation has the desired properties of working for points on the boundary and is interpretable. However, while such a transformation might account for the overall change, there still needs to be a model for the differences between the overall change applied to the data and the data observed in the next year. To this end, we explore in chapter 2 the additive Dirichlet error, exponential rotation error, and rotational geodesic error models to see how we can perform inference with these models. We will also discuss if these models are sufficiently rich, but still have the aforementioned properties.

As demonstrated in chapter 3, we realized that assuming observations moved in an uniform direction after applying the affine transformation was incorrect for the dataset. Instead, this shift appears dependent on the current income proportions and not on the neighborhood’s physical location. Further, the random direction is "noisy" because while there may be a clearly preferred direction, not all observations nearby move in this direction. One additional challenge is that there appears to be multiple patterns in the random direction. In other words, there is "heterogenity" in the random direction pattern.

Inspired by ideas from directional statistics [45], we developed the spatially varying von Mises component mixture model (SvM-c) and spatially varying mixing probability for von Mises distributions model (SvM-p). At a high level, the former assumes that proportions close to each other are likely to move in the same average direction whereas the latter assumes that proportions close to each other have higher chance of following the same pattern of movement. In addition to providing more technical details on these models, chapter 3 discusses some statistical properties of these models and how to sample from these models. Both are helpful to practitioners because information from the former section can help them set priors. We then conclude the chapter with a simulation study and apply the models to the random directions extracted from our data set.

When these ideas are extended to higher dimensions, it is possible that the entire set of observations does not follow the same trend. Instead, depending on the sub-regions, certain observations might move in certain directions. These sub-regions might then be represented as simplices of smaller ambient dimensions. To make this representation even more parsimonious, we might assume that vertices are shared across simplices. In particular, it might be possible to organize the vertices into a tree such that the root node is shared by all simplices whereas the leaf nodes are shared by a simplex. As a result, it might be desirable to discover these tree-based mixture of probability
simplices. While this was not the original motivation for chapter 4, it is a potential use of the ideas discussed in that chapter.

Inspired by the nested Chinese restaurant process and the nested hierarchical Dirichlet process, we wanted to develop theory for tree-based mixtures of probability simplices [11, 54]. This is challenging because in addition to the tree-based mixture being a mixture, the admixture model is already an extension of the mixture model because each observation follows a different proportion of the mixing distributions. Still, it is worthwhile to understand these models theoretically. Practitioners might wonder if it is possible to identify the distribution or the tree despite their complexity. Such information might give further confidence in the significance of the estimated structure. Further, the development of theory might lead to better methodology. For instance, while it is a simpler model, ideas from Nguyen’s paper on the contraction of polytopes were used to develop geometric algorithms [53, 74, 75, 76]. Indeed, chapter 4 establishes posterior consistency in distribution and structure. Using these insights, we expand on existing geometric algorithms and suggest new ways to sample from this type of model.

Reflecting this incentive, we use these techniques in a simulation study and apply them to analyze articles from the New York Times Corpus. This annotated set contains over 1.8 million documents from 1987 to 2007 [63]. However, due to memory issues, we examined a subset of articles from 2000. First, we removed all English stopwords from the SMART information retrieval system and from the stopwords function with "en" as its argument of the R tm package [22]. We further reduced our vocabulary size by examining the 1000 most common words across all articles from 2000. Then, we kept 30% of the articles that had more than 100 words of these words from each month. This left the corpus with 12,779 articles to analyze.

We conclude this thesis by discussing future directions to explore. As discussed earlier, one primary direction might be linking up the work in chapter 4 to chapters 2 and 3. This is not straightforward because the work in the next two chapters need to be linked up as well and extended to higher dimensions. Another direction might be computation. The models introduced might be difficult and computationally expensive to fit, particularly as the sample size and dimension of the data grow. Finally, one last direction might be in the further development of theory.
CHAPTER 2

Simplex Movement

2.1 Introduction

Modeling how data that lie on the simplex change can be challenging. The most immediate difficulty is the simplex constraint imposed by the proportions because they must sum up to one [57]. For this constraint to be maintained, a change in one proportion necessarily involves a change in the other proportions. As discussed earlier, if this is not handled correctly, the model might detect what Pearson called "spurious correlation" [58]. Another difficulty may arise if any of the proportions are zero. The log ratio transforms mentioned in the introduction become undefined if any of the proportions are zero.

To deal with this difficulty, we might model how the data changes from one time point to the next as movements of points within the simplex. There might be randomness in this movement. Despite the potential challenge in modeling movements that respect the simplicial constraint and are potentially "noisy", such an approach has several fundamental advantages. It would avoid picking up "spurious correlation" because all components of an observation would be dealt with simultaneously. Moreover, while the range of movements is restricted, boundary points could be treated in a similar framework as interior points.

To this end, our approach is to combine an affine transformation with a sufficiently rich "error" model. This transformation is a modified affine combination between two points. By applying this transformation to all data points at a time and one point in the simplex, we can uncover the overall change at one time point. However, there still might be differences between the transformed data and the data observed at the next time point. To model this, we need an "error" or "noise" model that is robust, valid on the boundary, and can move data from anywhere to anywhere in the simplex, including the boundary. In particular, we will see if the additive Dirichlet, exponential rotational, and rotational geodesic error models are suitable candidates. This exploration and understanding how to fit them is the primary focus of this chapter because despite the importance of the "noise" model having all three qualities, it is difficult to come up with such a model.
From this exploration, we find that the rotational geodesic error model best suits our purpose. Consequently, we use this combination to analyze the Los Angeles County income proportion data set. Doing so reveals trends that are consistent with larger macroeconomic ones. In addition, because our models are interpretable, we can characterize how strong this trend is.

The rest of the chapter is organized as follows. First, we define the affine transformation and discuss a few properties. Next, we describe our initial models, the additive Dirichlet noise and exponential rotational error model, and the potential problems associated with that model. We then introduce the rotational geodesic error model and describe how to compute the geodesic used by that model for the two dimensional case. These descriptions include how to fit such a model. Finally, we discuss results from simulation studies and the analysis of the Los Angeles County income proportion data set.

### 2.2 Affine transformation

As stated earlier, to model the transformation in the simplex data, we will combine an affine transformation with a sufficiently rich error model. By affine transformation, we mean the following.

**Definition 2.2.1.** For \(x \in \Delta^D\), given a coordinate vector, \(e_i\), for some \(i \in 1, 2, \ldots, D + 1\) and \(\eta \in (-\infty, 1]\),

- If \(x_i \geq \frac{-\eta}{1-\eta}\), let
  \[
  T_{\eta,e_i}(x) = (1 - \eta)x + \eta e_i.
  \]

- Otherwise, let
  \[
  T_{\eta,e_i}(x) = \left(1 - \frac{x_i}{x_i - 1}\right)x + \frac{x_i}{x_i - 1} e_i.
  \]

Denote \(T_{\eta,e_i}(x)\) as the **affine transformation**.
This transformation has the desired properties stated earlier. First, as we can see in Figure 2.1, the transformation is interpretable. If \( \eta > 0 \), then the data is being pulled toward some category whereas if \( \eta < 0 \), then the data is being pushed away from some category. Based on the particular category, we can assign meaning to the transformation. In addition, the affine transformation is crucially simplex preserving.

**Lemma 2.2.1.** For \( \mathbf{x} \in \Delta^D \), \( \eta \in (-\infty, 1] \), a coordinate vector, \( \mathbf{e}_i \), for some \( i \in 1, 2, \ldots, D + 1 \), \( T_{\eta, \mathbf{e}_i}(\mathbf{x}) \in \Delta^D \).

**Proof.** We have that

\[
\sum T_{\eta, \mathbf{e}_i}(x) = (1 - \eta) \sum x + \eta \sum \mathbf{e}_i = (1 - \eta) + \eta = 1.
\]

As a result, if \( T_{\eta, \mathbf{e}_i}(\mathbf{x}) j \geq 0 \forall j \in 1, 2, \ldots, d \), then \( T_{\eta, \mathbf{e}_i}(\mathbf{x}) j \in [0, 1] \forall j \). In particular, as \( (\mathbf{e}_i) j = 0 \) \( \forall j \neq i \), \( \mathbf{x} \in \Delta^n \), and \( 1 - \eta \geq 0 \), we only need to check \( T_{\eta, \mathbf{e}_i}(\mathbf{x}) i \geq 0 \). Then,

\[
T_{\eta, \mathbf{e}_i}(\mathbf{x}) i \geq 0 \\
(1 - \eta)x_i + \eta \geq 0 \\
x_i - \eta(x_i - 1) \geq 0 \\
x_i \geq \eta(x_i - 1) \\
\frac{x_i}{x_i - 1} \leq \eta.
\]

Because this is true by Definition 2.2.1, the affine transformation is simplex preserving.

As the proof doesn’t depend on the location of \( \mathbf{x} \), the affine transformation remains valid and simplex preserving even if \( \mathbf{x} \) is on the boundary.

We can then generalize the affine transformation in the following way.

**Definition 2.2.2.** For \( \mathbf{x} \in \Delta^d \), given \( \mathbf{v} \in \Delta^d \) and \( \eta \in (-\infty, \infty) \),

- If \( \forall i \in 1, 2, \ldots, d \) and for \( j \in 1, 2, \ldots, d \), \( x_i \leq \min_j \frac{-\eta}{1 - \eta} v_j \) for \( \eta > 1 \) or \( x_i \geq \max_j \frac{-\eta}{1 - \eta} v_j \) for \( \eta \leq 1 \), let

  \[
  T_{\eta, \mathbf{v}}(\mathbf{x}) = (1 - \eta)\mathbf{x} + \eta\mathbf{v}.
  \]

- Otherwise, for \( j \in 1, 2, \ldots, d \), \( \exists i \in 1, 2, \ldots, d \) such that \( \frac{x_i}{x_i - v_i} = \min_j \frac{-\eta}{1 - \eta} v_j \) for \( \eta > 1 \) or \( \frac{x_i}{x_i - v_i} = \max_j \frac{-\eta}{1 - \eta} v_j \) for \( \eta \leq 1 \). Let

  \[
  T_{\eta, \mathbf{v}}(\mathbf{x}) = \left(1 - \frac{x_i}{x_i - v_i}\right)\mathbf{x} + \frac{x_i}{x_i - v_i} \mathbf{v}.
  \]
Denote $T_{\eta,v}(x)$ as the **general affine transformation**.

This generalization is still simplex preserving.

**Lemma 2.2.2.** For $x, v \in \Delta^D$, $\eta \in (-\infty, 1]$, for some $i \in 1, 2, \ldots D + 1$, $T_{\eta,v}(x) \in \Delta^D$.

**Proof.** As in the previous proof, $\sum T_{\eta,v}(x) = 1$. Hence, it suffices to check that $(1 - \eta)x_j + \eta v_j \geq 0$ $\forall j \in 1, 2, \ldots, d$. Then, $\forall j \in 1, 2, \ldots, D + 1$,

\[
\begin{aligned}
(1 - \eta)x_j + \eta v_j &\geq 0 \\
x_j - \eta(x_j - v_j) &\geq 0 \\
x_j &\geq \eta(x_j - v_j)
\end{aligned}
\]

If $x_j - v_j = 0$, then as $x \in \Delta^D$, this condition is satisfied. If this is true $\forall j$, then $x = v$. We can set the result of the transformation to be $x$ and this transformation is always valid.

On the other hand, if $\exists j \in 1, 2, \ldots, D$ such that $x_j > v_j$, then because $\sum_j x_j = \sum_j v_j = 1$, $\exists j'$ such that $x_{j'} < v_{j'}$. The same argument holds if $x_j < v_j$. The condition listed in the definition ensures that we satisfy the inequality constraints. 

We introduced two simplex preserving transformations that are also interpretable. To better understand these transformations, we can further examine their mathematical properties. If we do so, we find that the affine transformation and the general affine transformation are associative, but not commutative. However, because we will be largely interested in only identifying one such transformation, we defer this discussion to the appendix.

## 2.3 Error Models

In the previous section, we introduced the affine transformation and the general affine transformation. Such a transformation makes sense to explicate a global trend. However, there might be differences between the transformed data from one time point and the observed data at another. The "noise" or "error" model that we define to explicate this error might make certain affine transformations more likely. As a result, how we define the model that explains this error affects how we will do inference. To that end, we propose two different classes of models and explain how to perform inference on them: the "additive" noise models and "rotational" noise models.

At a high level, the "additive" error models can be explained generatively. After an observation has been transformed with the affine or general affine transformation, these models return a convex combination of the transformed observation and a point drawn from a distribution on the simplex. In other words, we "add" noise to the observation. The example we will discuss in this chapter...
is the additive Dirichlet error model. We also tried to make the model richer by using multiple affine transformation on the observation. However, we ran into difficulties checking whether the transformed data lie in the simplex or not. As a result, we left discussion of this model to Appendix A.2.3.

On the other hand, the "rotational" error models decompose the "error" into a direction and a movement. It then assigns probability to both parts. For this chapter, the direction will be uniformly distributed. As a result, the focus is on assigning probability to the movement. The two examples discussed are the exponential rotational error model and the rotational geodesic error model.

### 2.3.1 Additive Dirichlet error model

Because we are trying to transform data on a simplex, a natural choice for an error model is the Dirichlet distribution. This is a popular prior distribution for parameters that lie on a simplex. This is used in the additive Dirichlet error model. Given \( \alpha \in \mathbb{R}^{(D+1)+} \) and a coordinate vector \( e_i, i = 1, 2, \ldots, D+1 \), it can easily be described in a generative manner for a data set, \( x_1, x_2, \ldots, x_N \in \Delta^D \), and parameters, \( \eta, \gamma \in (-\infty, 1] \):

1. Let \( z_n = T_{\eta, e_i}(x_n) \) for \( n = 1, 2, \ldots, N \).
2. Draw \( \epsilon_n \sim \text{Dir}(\alpha) \) for \( n = 1, 2, \ldots, N \).
3. For \( \gamma \in [0, 1] \), let \( y_n = (1 - \gamma)z_n + \gamma \epsilon \) for \( n = 1, 2, \ldots, N \).

In essence, we are applying our affine transformation twice. First, we apply it with a fixed coordinate vector. Then, we apply it to the transformed data with a point randomly drawn from a Dirichlet distribution.

Because of this choice, \( y_n \) cannot lie on the boundary for non-degenerate \( \alpha \) or non-trivial \( \gamma \). However, if this condition is met, the Dirichlet distribution gives us structure that we can exploit. After all, using change of variables, the average log likelihood for \( y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n, \gamma, \eta \) is the following:

\[
\frac{1}{n} \left( \sum_{n=1}^{N} - \log(B(\alpha)) - \sum_{j=1}^{d} (\alpha_j - 1) (\log((y_n)_j - (1 - \gamma)T_{\eta, e_i}(x_n)_j) - \log(\gamma)) - \log(\gamma) \right),
\]

(2.1)

where \( B(\alpha) = \frac{\prod_{j=1}^{d} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{d} \alpha_j)} \). However, it also requires that we check whether the affine transformed data with the additive Dirichlet error model is in the simplex for the parameters. We denote this as checking whether the simplicial constraints are satisfied or not. While the proofs are in Appendix A.2, the gradients of the log likelihood and simplicial constraints are the following for \( \eta \) and \( \gamma \).
It is relatively straightforward to derive the gradient and simplicial constraint for \( \gamma \) given \( \eta \).

**Lemma 2.3.1.** Given \( \eta, \alpha \), and a coordinate vector \( e_i \) for some \( i = 1, 2, \ldots, D + 1 \), the gradient for \( \gamma \) based on the average log likelihood is:

\[
\frac{d}{d\gamma} \text{ll}(y_1, y_2, \ldots, y_N \mid x_1, x_2, \ldots, x_N, \gamma, \eta) = \frac{1}{n} \left( \sum_{n=1}^{N} \sum_{j=1}^{d} \frac{(\alpha_j - 1)T_{\eta,e_i}(x_n)_j}{y_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j} \right) - \sum_{j=1}^{d} \frac{\alpha_j}{\gamma}.
\]

**Lemma 2.3.2.** The simplicial constraints are satisfied if

\[
\gamma \geq \max_{n,j} \left( -\frac{(y_n)_j - T_{\eta,e_i}(x_n)_j}{T_{\eta,e_i}(x_n)_j} \right).
\]

On the other hand, finding the gradient for \( \eta \) in this model is a bit more involved. We need the following definition to see why.

**Definition 2.3.1.** For a given \( x \in \Delta^D \), \( \eta \in (-\infty, 0] \), and coordinate vector \( e_i \) for \( i = 1, 2, \ldots, D+1 \), say \( x \in \Delta^D \) is active if \( \eta \geq \frac{-(x)_i}{1 - (x)_i} \) and inactive otherwise.

This is important because for the \( \eta \) that \( x_n \) is inactive, the log likelihood is constant so the gradient with respect to \( \eta \) will be zero. In other words, depending on \( \eta \), certain \( x_n \) cannot provide gradient information. Then, for \( n = 1, 2, \ldots, N \), let \( \xi_n = 1 \) denote that \( x_n \) is active and \( \xi_n = 0 \) denote that it isn’t. This leads us to the following lemma.

**Lemma 2.3.3.** Given \( \gamma, \alpha \), and a coordinate vector \( e_i \) for some \( i = 1, 2, \ldots, D + 1 \), the gradient for \( \eta \) based on the average log likelihood is:

\[
\frac{d}{d\eta} \text{ll}(y_1, y_2, \ldots, y_N \mid x_1, x_2, \ldots, x_N \gamma, \eta_i) = \frac{1}{n} \left( \sum_{n=1}^{N} \sum_{j=1}^{d} \frac{(\alpha_j - 1)(1 - \gamma)((x_n)_j - (e_i)_j)}{y_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j} \mathbb{1}(\xi_n = 1) \right).
\]

We again need to check that the simplicial constraints are met for active \( x_n \).

**Lemma 2.3.4.** For active \( x_n \), the simplicial constraint is satisfied if

\[
\max_{j \neq i} \left( -\frac{(y_n)_j - (1 - \gamma)(x_n)_j}{(1 - \gamma)(x_n)_j} \right) \leq \eta \leq \left( \frac{(y_n)_i - (1 - \gamma)(x_n)_i}{(1 - \gamma)(1 - (x_n)_i)} \right).
\]
However, as seen in the following lemma, inactive \( x_n \) can still affect whether the simplicial constraints are met.

**Lemma 2.3.5.** For inactive \( x_n \) and its corresponding \( y_n \) and \( \eta \leq 0 \), if

\[
\max_{n,j \neq 1} \left( \frac{T_{\eta}(x_n)}{y_n} \right) \leq \frac{1}{1 - \gamma},
\]

then the simplicial constraint is satisfied.

Based on these lemmas, given a direction \( e_i \) and \( \alpha \), the maximum likelihood estimate of \( \eta \) and \( \gamma \) can be found using the following procedure:

- **\( \gamma | \eta \)**: Run gradient ascent using Lemma 2.3.1 while checking that Lemma 2.3.2 is satisfied.
- **\( \eta | \gamma \)**: Run the following approach.

1. Let \( b_n = \frac{-(x_n)}{1-(x_n)} \). Sort \( b_n \) to create a grid \((b_{(1)}, b_{(2)}, \ldots, b_{(n)})\).
2. For \((b_{(n)}, b_{(n+1)})\),
   (a) Check that Lemma 2.3.5 holds for all inactive \( x_n \)'s and their \( y_n \)'s. If not, skip over the interval.
   (b) The interval must intersect with the following interval defined in Lemma 2.3.4 using active \( x_n \)'s and their \( y_n \)'s. If the intersection results in an empty interval, skip over the interval.
   (c) Using only the active \( x_n \)'s and their \( y_n \)'s, the gradient on the intersected interval can be found using Lemma 2.3.3.

We now make a few comments about this procedure. While our lemmas led naturally to maximum likelihood estimates, they also give us insight about the structure of the likelihood and thus might be useful when attempting to perform Bayesian inference. In addition, this model requires that we check whether the parameters satisfy the simplicial constraint. A natural question is whether this affects inference because the area for which the simplicial constraint is satisfied might be discontinuous or the true value might be close to regions for which the constraint is not met. However, due to the math not being straightforward, we answer this question using simulation. The results can be seen in Section 2.4.1.
2.3.2 Exponential rotational error model

The next model is part of a class of rotational error models. Because an observation is restricted in how far it can move in a given direction due to the simplicial constraints, the truncated exponential distribution provides one way to assign probability to the movement. This idea motivates the exponential rotational error model. Given some \( \lambda \in \mathbb{R} \), a distance function \( d(\cdot, \cdot) \), and a coordinate vector \( e_i \), the model is described in the following generative manner for parameter, \( \eta \in (-\infty, 1] \), and data, \( x_1, x_2, \ldots, x_N \in \Delta^D \):

1. Let \( z_n = T_{\eta, e_i}(x_n) \) for \( n = 1, 2, \ldots, N \).
2. Draw a direction, \( s_n \), uniformly from the \( D - 1 \) dimension sphere, \( S^{D-1} \). Let \( u_{n,i} = s_i \) for \( i \in 1, 2, \ldots D \) and \( u_{D+1} = 1 - \sum_{j=1}^{D} u_{n,j} \).
3. Find \( z_n^{bdy} \) such that \( z_n^{bdy} = z_n + \eta' u_n \) for some \( \eta' \in \mathbb{R}^+ \) such that \( \prod_j z_j^{bdy} = 0 \), and \( z_j^{bdy} \geq 0 \) \( \forall j \in 1, 2, \ldots D \).
4. Let \( d_{\max}^n = d(z_n, z_n^{bdy}) \). For \( \lambda > 0 \), draw \( c_n \sim \text{Exp}_{[0, d_{\max}]}(\lambda) \). Set \( y_n = z_n + \eta'' u \) for some \( \eta'' \) such that \( d(y_n, z_n) = c_n \).

We now make a few comments about this model. First, based on this description, choosing the distance function is important. While there are a few choices, the one that we use is the simplex Hellinger distance function. For \( x_n, x'_n, n, n' \in 1, 2, \ldots, N \) the simplex Hellinger distance, \( d_H(\cdot, \cdot) \), is defined below:

\[
d_H(x_n, x'_n) = \frac{1}{\sqrt{2}} \sqrt{\sum_i \left( \sqrt{x_{n,i}} - \sqrt{x'_{n,i}} \right)^2}.
\] (2.2)

This choice is motivated by potential problems with the \( L_2 \) distance. In particular, if there are many points near the boundaries, then the model prefers an \( \eta \) that sends points to the boundary. The \( L_2 \) distance between two points in the simplex does not change with respect to the location of the points. On the other hand, two points that have the same \( L_2 \) distance are closer according to the the simplex Hellinger distance if they are in the center compared to if they are near the boundary. In addition, the simplex Hellinger distance remains valid for points that lie on the boundary.

Next, another question is whether for a given \( z_n \), \( y_n \) remains in the simplex. Clearly, this noise model ensures that \( y_n \) remains in the simplex for any \( z_n \) in the interior of the simplex. For \( z_n \) that lie on the boundary, we modify the generative process. If the direction picked is away from the simplex, then the transformed point stays with probability 1. Otherwise, it uses a truncated exponential function based on a distance function to determine where it ends up in a given direction. With this simple idea, the simplicial constraint is satisfied.
Finally, we might wonder how to perform inference on this model. At a high level, a simple optimization approach to find $\eta$ would involve minimizing the distance between the affine transformed source data and the observed transformed data. To prevent $\eta$ from being too small or too large, we might add a penalization term. In other words, we might want to minimize the following:

$$
\sum_n \lambda \| y_n - T_{\eta,e_i}(x_n) \|_d - (a - 1) \log(1 - \eta) + b(1 - \eta).
$$

(2.3)

Here, the norm is with respect to some distance function, $d(\cdot, \cdot)$, such as the Hellinger distance on the simplex. This function also is a negative unnormalized log posterior because the first part is the negative kernel of an exponential distribution and the latter part is the negative kernel of a gamma distribution. Indeed, we might consider the first part the likelihood and the latter part the prior.

As a result, using all the ideas discussed previously, we can determine the MCMC acceptance probability in the following way. For an observation $x_n, y_n$ for $x = 1, 2, \ldots, N$, we have that

$$
p(x_n, y_n | \eta, \lambda) = \frac{1}{\text{SA}(S^{D-1})^{\text{exp}[0,a \cdot d_{\text{max}}]}(c_n | \lambda)} \left[ \prod_{c_n \neq 0} \left( \int_{s''_n \in S^{D-1}} \prod \left( u''_n + z_n \in \Delta^D \right) \right) \right].
$$

(2.4)

Here, $\text{SA}(\cdot)$ denotes the surface area and $u''_n$ is the $s''_n$ transformed according to the generative model. Then, for a given $a, b \in \mathbb{R}$, the MCMC acceptance probability is the following for a proposal distribution function, $g(\cdot | \cdot)$:

$$
\min \left( \frac{\prod_n p(x_n, y_n | \eta', \lambda) \Gamma(1 - \eta' | a, b) g(\eta | \eta')} {\prod_n p(x_n, y_n | \eta, \lambda) \Gamma(1 - \eta | a, b) g(\eta' | \eta')} , 1 \right).
$$

(2.5)

### 2.3.3 Rotational geodesic error model

For the exponential rotational error model, we described sending the transformed point in a given direction according to the truncated exponential distribution. What we elided in that description is that while we are sending the point according to the L2 geodesic for a given direction, we are essentially measuring it according to some other distance function. However, these other distance functions define their own geodesics on the simplex space. It would be more consistent to use a given distance function and the geodesics it defines.

This motivates us to develop the rotational geodesic error model. As before, the distance function we consider is the simplex Hellinger distance. By definition, the simplex Hellinger distance is the L2 distance on the positive orthant. As a result, the geodesic that the distance defines are the great circles. Then, given coordinate vector, $e_i$, the model can be defined in the following generative manner for parameters, $\eta \in (-\infty, 1], (\psi_0, \psi_1, \psi_2) \in \Delta^2$, and $b \in \mathbb{R}^+$; and data, $x_1, x_2, \ldots, X_N \in \Delta^D$:  

\[16\]
1. Let $z_n = T_{\eta,e_i}(x_n)$ for $n = 1, 2, \ldots, N$.

2. Draw a point $u_n$ in an uniform random direction $\phi$ away from $z_n$.

3. Define $g(t)$ to be the simplex Hellinger distance geodesic in the direction of $\phi$ such that $g(0) = z_n$, $g(t') = u$ for some $t' \in \mathbb{R}^+$, $g(1) \geq 0$, and $\prod_j g(1)_j = 0$.

4. Draw $\zeta_n \sim \text{Cat}(\psi_0, \psi_1, \psi_2)$. Then,
   - If $\zeta_n = 0$, set $y_n = g(0)$.
   - If $\zeta_n = 1$, set $y_n = g(1)$.
   - Otherwise, draw $c_n \sim \text{Beta}(1, b)$. Set $y_n = g(t)$ such that $d_H(y_n, z_n) = d_H(g(0), g(1)) = c_n$ for some $t \in (0, 1)$.

To perform inference, we need to compute the geodesic. We begin by discussing how to do so for $\Delta^2$. For ease of presentation, we drop the subscript $n$ and move the proofs to Section A.3.1 in the appendix. Notice that if $x \in \Delta^3$, we can use spherical coordinates to represent $\sqrt{x}$ as $(\sin(\theta) \cos(\phi), \sin(\theta) \sin(\phi), \cos(\theta))$ where $\phi \in [0, \frac{\pi}{2}]$ and $\theta \in [0, \frac{\pi}{2}]$. We restrict $\phi$ and $\theta$ in this manner because the simplex Hellinger distance is the L2 distance on the positive orthant.

Rewriting the arc-length formula in spherical coordinates and solving the Euler equation, we get that the geodesic path is

$$\cot(\theta) = a \cos(\phi - \phi_0), \quad (2.6)$$

where $a \in \mathbb{R}^+$ and $\phi_0 \in (-\frac{\pi}{2}, \frac{\pi}{2})$ are constants that we have to solve using our start points [34]. These constants can be found by the following lemma.

**Lemma 2.3.6.** If $z$ and $y$ are our two points on the simplex Hellinger distance geodesic, then

$$\phi_0 = \tan^{-1}\left(\frac{z_3 y_1 - z_1 y_3}{z_2 y_3 - z_3 y_2}\right),$$

$$a = \frac{z_3}{z_1 \cos(\phi_0) + z_2 \sin(\phi_0)}.$$

After calculating $a$ and $\phi_0$ and assuming $z \neq y$, there are two ways to use this geodesic equation to calculate the endpoints. We can do so due to Hopf-Rinow. One approach is the following:

1. Find the spherical coordinates for $z$, i.e. $\phi_z$, and $\theta_z$, and $y$, i.e. $\phi_y$, and $\theta_y$. 


2. Check whether $\phi = 0$, $\phi = \frac{\pi}{2}$, or $\phi = \frac{\pi}{2} + \phi_0$ satisfies our constraints because the first two represent $z_1 = 0$ and $z_2 = 0$ respectively in spherical coordinates whereas the third implies that $z_3 = 0$ due to the geodesic path in (2.6). Then, $\theta$ can be computed for these $\phi$ using (2.6).

3. If we have multiple choices, we can check whether $\phi$ is increasing between $\phi_z$ and $\phi_y$. If it is, we choose the larger valid $\phi$. If $\phi_z = \phi_y$, then we can use the lemma below.

**Lemma 2.3.7.** If $\phi_z = \phi_y$ and $z \neq y$, then the geodesic endpoint is $(0, 0, 1)$ if $\theta_z > \theta_y$ and $(\cos(\phi_z), \sin(\phi_z), 0)$ if $\theta_z < \theta_y$.

Another approach is to use $z_3 = a(z_1 \cos \phi_0 + z_2 \sin \phi_0)$ and $z_1^2 + z_2^2 + z_3^2 = 1$ to calculate the end points. Then,

- $z_1 = 0 : \begin{align*} z_1' &= \frac{1}{\sqrt{a^2 \sin^2 \phi_0 + 1}}, \\
  z_2' &= \frac{a \sin \phi_0}{\sqrt{a^2 \sin^2 \phi_0 + 1}}; \\
  z_3' &= \frac{a \cos \phi_0}{\sqrt{a^2 \cos^2 \phi_0 + 1}};
\end{align*}$

- $z_2 = 0 : \begin{align*} z_1' &= -\sin(\phi_0), \\
  z_2' &= \cos(\phi_0).
\end{align*}$

We again check which is a valid endpoint. If there are multiple valid constraints, then we should check whether $z_3 > y_3$. If so, we should pick the smaller of the two valid values. Otherwise, we should pick the larger of the valid values.

These ideas can then be generalized to higher dimensions. However, because the data set that we ultimately analyze is in $\Delta^2$, we leave the details to Section A.3.3 in the appendix. Instead, we now discuss how to MCMC sampling. For the model, we use the following priors:

$$1 - \eta \sim \Gamma(\cdot \mid \alpha_{\eta}, \beta_{\eta}),$$

$$b - 1 \sim \Gamma(\cdot \mid \alpha_b, \beta_b),$$

$$(\psi_0, \psi_1, \psi_2) \sim \text{Dir}(\cdot \mid \alpha_\psi).$$

Then, we have the following Gibbs sampler:

- To sample $\eta, \zeta_1, \zeta_2, \ldots, \zeta_n$ | $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, \psi_0, \psi_1, \psi_2, b$, first check that $\eta > \eta_{\text{min}}$, where $\eta_{\text{min}}$ satisfies the following inequality:

$$\min_n \max_j \frac{(x_n)_j}{\max_j \min_{i,(x_i)_j-(e_i)_j<0} (x_n)_j - (e_i)_j} \leq \eta_{\text{min}} \leq \max_n \min_j \frac{(x_n)_j}{\max_j \min_{i,(x_i)_j-(e_i)_j>0} (x_n)_j - (e_i)_j}.$$ (2.7)
Notice that \( c_n \) sets \( \zeta_n \) for \( n = 1, 2, \ldots, N \). Then, the likelihood is given below:

\[
p(x_n, y_n \mid \eta, \zeta_n, \psi_0, \psi_1, \psi_2, b) = \prod_n \left( \psi_0 \int_{\Delta^D} \mathbb{1} \left( g(t > 0) \notin \Delta^D \right) \right)^{\mathbb{1}(\zeta_n=0)} \left( \frac{\psi_1}{\text{SA}(S^{D-1})} \right)^{\mathbb{1}(\zeta_n=1)} \left( \frac{\psi_2}{\text{SA}(S^{D-1})} \right)^{\mathbb{1}(\zeta_n=2)} \text{Beta}(c_n \mid 1, b'). \tag{2.8}
\]

As a result, given an a priori proposal distribution \( g_{\eta}(. \mid .) \), the acceptance probability is the following:

\[
\frac{\left( \prod_n p(x_n, y_n \mid \eta', \zeta_n, \psi_0, \psi_1, \psi_2, b) \right) \Gamma(1 - \eta' \mid \alpha_n, \beta_n) g_{\eta}(\eta \mid \eta')}{\left( \prod_n p(x_n, y_n \mid \eta', \zeta_n, \psi_0, \psi_1, \psi_2, b) \right) \Gamma(1 - \eta \mid \alpha_n, \beta_n) g_{\eta}(\eta' \mid \eta)}. \tag{2.9}
\]

- Draw \( \psi_0, \psi_1, \psi_2 \mid x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, \eta, \zeta_1, \zeta_2, \ldots, \zeta_n, b \) from \( \text{Dir}(\alpha_\psi) \), where \( (\alpha_\psi)_i = (\alpha_\psi)_i + \sum_n \mathbb{I}(\zeta_n = i) \) for \( i = 1, 2, 3 \).
- Sample \( b \mid x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, \eta, \zeta_1, \zeta_2, \ldots, \zeta_n, \psi_0, \psi_1, \psi_2 \) with the following acceptance probability given a proposal distribution \( g_{\eta}(\cdot \mid \cdot) \):

\[
\frac{\left( \prod_n \text{Beta}(c_n \mid 1, b') \mathbb{1}(\zeta_n=2) \right) \Gamma(b' - 1 \mid \alpha_b, \beta_b) g_{b}(b \mid b')}{\left( \prod_n \text{Beta}(c_n \mid 1, b) \mathbb{1}(\zeta_n=2) \right) \Gamma(b - 1 \mid \alpha_b, \beta_b) g_{b}(b' \mid b')). \tag{2.10}
\]

As discussed in the next section, we also are interested in using this noise model with the general affine transformation. If \( v \) denotes the vector used in the general affine transformation, the priors are updated to be the following:

\[
\eta \sim N(. \mid \mu_\eta, \sigma_\eta^2), \quad v \sim \text{Dir}(\cdot \mid \alpha_v).
\]

Here, the normal distribution is truncated. However, we do not need to compute the truncation constant because the constant is the same regardless of \( \eta \). Then, while this implicitly updates all steps in the Gibbs sampler, there are explicit updates for the first step. The inequality check in (2.7) is changed and uses \( v \) instead of \( e_i \). Further, while the likelihood in (2.8) is inherently updated, the acceptance probability in (2.9) is now the following for a given proposal function, \( g_{\eta}(\cdot \mid \cdot) \):

\[
\frac{\left( \prod_n p(x_n, y_n \mid \eta', \zeta_n, \psi_0, \psi_1, \psi_2, b, v) \right) N(\eta' \mid \mu_\eta, \sigma_\eta^2) \text{Dir}(v' \mid \alpha_v) g_\eta(\eta \mid \eta') g_v(v \mid v')}{\left( \prod_n p(x_n, y_n \mid \eta', \zeta_n, \psi_0, \psi_1, \psi_2, b, v) \right) N(\eta' \mid \mu_\eta, \sigma_\eta^2) \text{Dir}(v \mid \alpha_v) g_\eta(\eta' \mid \eta) g_v(v' \mid v)}. \tag{2.11}
\]
Figure 2.2: Plots showing affine transformed data on the left and affine transformed data with the additive Dirichlet error model on the right ($\gamma = 0.1$, $\alpha = 2$). Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$.

2.4 Simulation studies

In order to better understand our models, we ran simulation studies. In doing so, we wanted to understand whether it was possible to recover our model parameters and illuminate any potential shortcomings. The latter was particularly important as we attempted to decide which model to use.

2.4.1 Additive Dirichlet error model

For the additive Dirichlet error model, the parameters of interest were $\eta$ and $\gamma$. We generated 2000 data points from Dir(2.5, 1.25, 1.25). We then followed the generative model with $\eta = -0.2$ and $\eta = 0.1$, $\gamma = 0.1$, and $\alpha = (2, 2, 2)$. From Figure 2.2, we see that our additive Dirichlet error provides a pull back for points. It is particularly noticeable for $\eta = -0.2$ because we see points pulled from the boundary. On the other hand, for $\eta = 0.1$, the points are being pulled closer together through the additive Dirichlet error model by looking at the range of the x-axis.

These initial plots highlight one potential concern that the model expects the data to be close
to Dirichlet. As discussed earlier, the additive Dirichlet error model imposes certain boundary conditions because \( y_n \) cannot lie on the boundary for non-degenerate \( \alpha \) or non-trivial \( \gamma = 0 \). While this might be appropriate for some data sets, the data set we are interested in modeling has points on the boundary. This makes it hard to use the additive Dirichlet error model for our data.

Another potential concern is shown in Figure 2.3. While we can recover the values that we used to simulate the observed data using gradient ascent from \( \gamma = 0.9 \) and various values of \( \eta \), we see that the value is on the boundary of where the log likelihood is defined. In addition, we see the optimization path hugs the boundary as well. This is problematic because even though the region for which the simplicial constraint is met is continuous, we might end up with values for which the log likelihood is not defined if we are not careful. It further demonstrates how our likelihood is sensitive to the choice of parameters. Further, as we saw in the gradient calculations, how sensitive the likelihood is can change depending on \( x_n \) and \( y_n \).

Due to these two concerns, we did not apply this noise model to our actual data. Still, because it is a simple model and straightforward to fit, it might be useful for other data sets that do not have points that lie on the boundary.

### 2.4.2 Exponential rotational error model

For simulation, we generated 2000 data points from \( \text{Dir}(5, 2.5, 2.5) \). We then followed the generative model with \( \eta = -0.2 \) and 0.1 and \( \lambda = 10 \). From Figure 2.4, unlike the additive Dirichlet error model, we see that the exponential rotational error model appears to increase the spread of the simulated data. For \( \eta = 0.1 \), the affine transformation pulls points toward a coordinate vector and shrinks the simplex because by definition, any point on the opposite boundary is moved to 0.1. However, due to the exponential rotational error model, many are sent back toward the opposite
Figure 2.4: Plots showing affine transformed data on the left and affine transformed data with the exponential rotational error model, $\lambda = 10$, on the right. Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$. 
Figure 2.5: Plots of the log likelihood for $\eta = -0.2$ and $\eta = 0.1$ and the expected end points given a direction for different values of $\lambda$. The x-axis represents different values of $\eta$.

boundary. Similarly, for $\eta = -0.2$, the simulated data appears less concentrated after applying the exponential rotational error model.

Interestingly enough, as seen in Figure 2.5, the log likelihood for the simulated data is very concentrated around the $\eta$ used to generate the data. As a result, while the model with the right direction can correctly identify $\eta$, the credible intervals are very tight in both simulation and real data when we perform Bayesian inference on $\eta$ using Equations (2.4) and (2.5). We ran an independent MCMC sampler with the same distribution on $1 - \eta$ as the prior, i.e. a Gamma distribution with $a, b$ set to be 4 and 3 respectively. This is shown in Figure 2.6. This is problematic because the model is not robust. If we generate the data via another model, then as seen in Figure 2.6, the $\eta$ used to simulate the data is not estimated and the credible intervals are not much wider to reflect the uncertainty.

Figure 2.5 also includes another plot that might elucidate why the model struggles. To generate it, we took the point $(1/3, 1/3, 1/3)$, drew 1000 random directions, and plotted the expected point for the directions for different values of $\lambda$. Interestingly, as $\lambda$ gets smaller, the circle becomes larger and shaped more like the simplex. While this suggests that we should also sample for $\lambda$, the fact that the expected value isn’t near the simplex edge is concerning. Indeed, by L’Hôpital’s rule, it can be shown that the limit as $\lambda \to 0$, the truncated exponential becomes a uniform distribution. A uniform distribution still doesn’t expect points to be close to the boundary. While it might be difficult to recover, we can imagine a generative mechanism that is expected to send points close to the boundary. In addition, we might imagine a generative mechanism in which the location of the expected point changes depending on the direction and is not uniform. In fact, those are the models that the exponential rotational error model struggles with in Figure 2.6. It is not clear if the general affine transformation will solve these issues.
Figure 2.6: Plots showing the posterior mean and credible intervals for $\eta$. The left two plots show simulation results under different model when $\eta$ is known and shown by the yellow line. The right plot are the estimates for the data transformation from one year to the next connected for visualization purposes.

2.4.3 Rotational geodesic error model

To understand our model, we simulated the data in three dimensions using the ideas in Section A.3.2 in the appendix. Taking the same starting data that we used for the exponential rotational error model, we applied the generative model with $\psi_0 = 0.6$, $\psi_1 = 0.02$, $\psi_2 = 0.38$ and Beta(1, 3) for the proportion probability. We again used $\eta = 0.1$ and $\eta = -0.1$. Based on Figure 2.7, the noise model mostly preserves the results of the affine transformation while also spreading some of the data out and helping to recover the simplex. After all, in addition to seeing the affine transformed pattern in the noise plots, we see points that lie in the full range of the simplex.

However, when we look at the log likelihood plot in Figure 2.8, we again see that the log likelihood is heavily concentrated around the true value. This concentration is due to the probability of moving being smaller than the probability of matching the point exactly. It can then be further penalized because we multiply the smaller probability by the Beta distribution.

One potential way to reduce the spikiness of the log is by using the general affine transformation with some destination point $v \in \Delta^d$. It is possible that more uncertainty might smooth the likelihood surface. In addition, even if we assume $v$ is in the interior, the general affine transformation can still capture the effect of the affine transformation by taking $v$ to be very close to a coordinate vector. However, that might be a difficult case to capture because it is so close to the boundary. As such, to fully understand the general affine transformation, we simulated cases in which $v = (1/3, 1/3, 1/3)$, $v_i = 1/2$ and the other coordinates are $1/4$, $v_i = 3/4$ and the other coordinates are $1/8$ for $i \in \{1, 2, 3\}$.

As we can see in Figure 2.9, there is still a significant spike at the $\eta$ used to generate the data. However, the rest of the likelihood surface is reasonably smooth and is higher in the area around the $\eta$ used to generate the data. Interestingly enough, if we look at the likelihood surface along the major axis, it is more strongly identified for $\eta$ negative versus $\eta$ positive. After all, the likelihood...
Figure 2.7: Plots showing affine transformed data on the left and affine transformed data with the rotational geodesic error model with $\psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38$ and $\text{Beta}(1, 3)$ for the proportion probability on the right. Top plots are for $\eta = -0.2$ and bottom plots are for $\eta = 0.1$.

Figure 2.8: Plots of the log likelihood for $\eta = -0.2$ and $\eta = 0.1$ under the rotational geodesic error model with $\psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38$ and $\text{Beta}(1, 3)$ for the proportion probability. The x-axis represents different values of $\eta$. 
Figure 2.9: Left plots show the likelihood contour for $v_1$, $v_2$, and $\eta \in [-.5, .5]$ with the log likelihood at $(1, 0, 0)$ removed whereas the right plots show the likelihood surface for $\eta \in [-.5, .5]$ along the line $v = (0, 1/2, 1/2)(1 - t) + (1, 0, 0)t$. The value used to simulate the data are $v = (1, 0, 0)$ and $\eta = -0.2$ for the top plots and $\eta = 0.1$ for the bottom plots. $\psi_0 = 0.6$, $\psi_1 = 0.02$, $\psi_2 = 0.38$ and Beta(1, 3) for the proportion probability were used to calculate the log probability.

surface around the $\eta$ used to simulate the data when $\eta$ is negative looks to be part of a mountaintop. On the other hand, the $\eta$ used to simulate the data appears part of a ridge when $\eta$ is positive. We can see this in the likelihood contour plots because there is only one hotspot when $\eta$ is negative whereas there appears to be a region of high probability connecting $(1, 0, 0)$ with the other end of the simplex.

This affected how we sampled the general affine case with rotational geodesic error model. We tried both a Gibbs approach to sample $\eta$ and $v$ and an MCMC approach to simultaneously sample both. Based on Figure 2.9, we ultimately went with the latter approach in order to better take advantage of the likelihood surface. As the contour likelihood plots for positive $\eta$ particularly shows, the high probability region extends across $\eta$ and $v$ instead of just one variable. A Gibbs approach might be slow in following this region because it can only move following $\eta$ or $v$. In addition, we initially tried a random walk for $\eta$ and $v$ so $v$ wouldn’t jump around the simplex and lose information about the surface. However, upon further investigation, we used the self-adjusting logit transform (SALT) proposal to sample for $v$ because we intend to apply our model to data in higher dimensions and the random walk proposal will not work as well in higher dimensions[19].
Figure 2.10: Trace plots of $\eta$ for $v_i = 1$ if $i$ is the true direction and $v_i = 0$ if $i$ is not the direction. The left plot shows results that start from $(1/3, 1/3, 1/3)$ and the right plot shows results from $(0.8, 0.1, 0.1)$. Yellow line indicates the true value, $\eta = 0.1$.

This also affected the posterior samples. We ran our MCMC approach on data generated from $\eta = 0.1$ and $\eta = -0.2$ and $v = (1/3, 1/3, 1/3)$ and $v = (1/2, 1/4, 1/4), (3/4, 1/8, 1/8)$, and $(1, 0, 0)$ in all directions. From doing so, we found that we can recover $\eta$ and $v$ in all but one case in simulation for $v = (1/3, 1/3, 1/3)$ and $v = (1/2, 1/4, 1/4)$, and $v = (3/4, 1/8, 1/8)$ in any direction. However, we did have difficulty recovering $\eta$ for the affine transformation case. This makes sense because as seen in Figure 2.9, if it starts at around $(1/3, 1/3, 1/3)$, it might be too far from the area of higher probability or it gets attracted to another area of high probability.

As such, one thing we tried was starting $v$ at $(0.8, 0.1, 0.1), (0.1, 0.8, 0.1), \text{and} (0.1, 0.1, 0.8)$. Figure 2.10 demonstrates that this both helped and hurt the sampler. By starting at $(0.8, 0.1, 0.1)$, the MCMC sampler can recover $\eta = 0.1$ for the case of $(1, 0, 0)$, but it completely misestimates $\eta$ for the cases of $(0, 1, 0)$ and $(0, 0, 1)$. This makes sense because starting at $(0.8, 0.1, 0.1)$ means the starting point is close to 1 in the first coordinate, but it is also close to 0 in the other coordinates. Still, it is useful to start at these different starting points because if they converge to the same value, that is a good sign that there is only one mode.

Another way to perhaps improve performance is increase the target area of the likelihood sampler is to also sample for $\psi_0, \psi_1, \psi_2$. We found that while the sampler gets the probability for $\psi_0, \psi_1, \psi_2$ wrong, it does not affect the recovery of $\eta$ and $v$. The sampler places a high probability of $\psi_2$ and a low probability on $\psi_0$ because it is hard to match the affine transformation exactly.

To understand why it is still able to recover $\eta$, we looked at of the log likelihoods for different $\psi_0$ and $\psi_2$ under the affine transformation and the same beta distribution for proportions. While this is a simplification of the model, we can still gain insights. For instance, the overall shapes of the log likelihoods in Figure 2.11 are similar. Instead, the only change appears to be the probability of the $\eta$ used to simulate the data. As a result, even if the middle figure in Figure 2.11 is closest to the sampled $\psi_0, \psi_1, \psi_2$ and the true value of $\eta$ has the lowest probability, the MCMC sampler still
Figure 2.11: Plots of the log likelihood under the affine transformation with $\eta = 0.1$ and the rotational geodesic error model with Beta(1, 3) for the proportion probability and different values of $\psi_0, \psi_1, \psi_2$. The x-axis represents different values of $\eta$.

While it is possible that $\psi_1$ can affect the log likelihood shape, the choice of the Beta distribution affects the probabilities assigned to proportions. The distribution might have more influence in controlling the overall shape of the log likelihood and whether the shape concentrates around the true value of $\eta$. As a result, if the proportions were assumed to be distributed around Beta(1, $b$), we might want to sample for $b, b \in \mathbb{R}, b > 1$ to increase the target area. Figure 2.12 provides further evidence supporting this idea. After running MCMC for $b \in [2.5, 3.5]$ and plotting the scaled posterior, we see that changing $b$ changes where the posterior peaks. Sampling for $b$ might smooth out the posterior. Further, despite this potential smoothing, accuracy might not be lost because the $\eta$ used to simulate the data is still included in the scaled posterior density plots.

Interestingly, when we do run a random walk starting at $b = 2$ with a standard deviation of 0.01, the value of $b$ is always increasing. The lower end of the final posterior 95% credible interval is usually around 5 and the upper end is around 7 after 2000 iterations. Even using a strong prior like Exp(10) does not affect this. Only using a smaller deviation of 0.0002 and starting at 2 causes the sampler to stay between $b = 2$ and $b = 3$ after the same number of iterations. In essence, it appears that as the model fits better to the data, it encourages larger $b$ to increase the probability that the proportions are small. This then encourages a tighter fit, creating a positive feedback cycle.

As a result, in simulation when we use a random walk starting at $b = 2$ with a standard deviation of 0.01, it performs as well in recovering the $\eta$ used to generate the data as if we hadn’t sampled for $b$ except that the credible intervals are much tighter. We then ran test cases on one data set to better understand the effect of $b$. While the model’s credible interval can still recover the $\eta$ used to simulate the data regardless of the choice of the starting location and random walk standard deviation, it seems that the interval of $b$ explored by the sampler past the burn-in period is important...
\( \psi_0 = 0.6, \psi_1 = 0.02, \psi_2 = 0.38 \)

\( \psi_0 = 0.01, \psi_1 = 0.02, \psi_2 = 0.97 \)

Figure 2.12: Plots of the scaled posterior for \( \eta = -0.2 \) and \( \eta = 0.1 \) and \( v = (1/3, 1/3, 1/3) \) after running MCMC for different values of \( b \in [2.5, 3.5] \). \( b \) is the parameter from Beta(1, \( b \)), which assigns probability to the proportions.

in determining the width of the credible intervals.

Based on this, the choice of the standard deviation and the starting location for \( b \) should vary. For instance, in simulation, we get reasonable estimates and credible intervals starting at \( b = 2 \) and a standard deviation of 0.002 because while the model is correct for the simulated data, we don’t want the MCMC sampler to be too certain. On the other hand, we might want to start at 2, but use a larger standard deviation for real data. The model might not be correct so we want the sampler to be more flexible in how confident it is. If we use a standard deviation that is too small and remain at smaller values of \( b \), this might encourage looser fits and lead to the model discovering multiple possible transformations instead of picking one.

### 2.5 Data analysis

Based on the discussion above, we ran 4 chains of the following blocked Gibbs sampler on our data for 5000 iterations, starting at \((1/3, 1/3, 1/3), (0.8, 0.1, 0.1), (0.1, 0.8, 0.1), (0.1, 0.1, 0.8)\). The priors used were Dir(\( \cdot \mid 1_3 \)), N(\( \cdot \mid 0, 1 \)), and \( \Gamma(\cdot \mid 8, 4) \) for \( v \), \( \eta \), and \( b \) respectively. Meanwhile, new values were proposed by the SALT sampler and random walks with standard deviation, 0.01.

When applied to the income proportions from Los Angeles County, our model tended to send as many points as possible to the boundary. To counteract this, we added a truncated exponential prior on the Hellinger distance. The resulting transformation better represented the transformation observed in the data set for that given year. Further, when we tried that same prior in simulation,
Figure 2.13: Plots showing the results of running the MCMC sampler and sampling all indices of $v$ before sampling $\eta$, starting at $(1/3, 1/3, 1/3)$ unless otherwise indicated.

The model performed as well as it did without the truncated exponential prior in recovering $\eta$ and $v$. However, the credible intervals were much tighter than before.

With this modification, the results of sampling all indices of $v$ and then $\eta$ can be seen in Figure 2.13. If we look at the trace plots for $\eta$, we see that for all but two years’ transformation, it looks like the sampler has converged and is mixing well. The estimate for $\eta$ and $v$ are fairly consistent across all chains. Further, if we apply the general affine transformation with these mean posterior estimates to our data, we can largely capture the change in the data with the exception of 2007-2008. Indeed, while the years with large positive $\eta$ shrink the space, the representation still largely looks similar to the data if we treat the shrunk space as the simplex. Not only that, there are interesting trends revealed in Table 2.1. For instance, the mean trend between 2001 and 2007 is a steady increase in the number of mortgages between $100,000 and $200,000, perhaps reflecting the housing bubble.
However, we see the effect of the 2007-2008 financial crisis because there is a repulsion away from the two highest income categories in 2008-2009. Unfortunately, except for the transformations between 1993 to 2002, the starting location can heavily affect the estimate of $\eta$. There might be multiple general affine transformations that can explain the observed changes.

On the other hand, the results of sampling an index of $v$ and then $\eta$ can be seen in Figure 2.14. Again, if we look at the trace plots for $\eta$, it looks like the sampler has converged and is mixing well. However, unlike before, it is the transformation from 1990 to 1991 that is giving the sampler difficulty. Further, it is interesting that the one chain that diverges from the other chain gives the same estimate of $\eta$ that the sampler discussed in the previous paragraph gives. It is particularly noteworthy because otherwise, as seen in Table 2.1, the two samplers give the same estimate for $\eta$ and $v$ after throwing out divergent chains. As a result, while there are multiple solutions, it is
We first introduced the affine transformation, which is a modified affine combination between two Angeles County income proportion data set with these models. The trends revealed agreed with the utility of interpretability.

Larger macroeconomic ones. The model additionally illustrated how strong they were, highlighting transformation parameter when the rotational geodesic error model was combined with the general these models. Out of the three models, the rotational geodesic error model seemed most promising, rotational, and the rotational geodesic error model. We also discussed how to perform inference on trend. To model this "noise", we introduced the additive Dirichlet error model, the exponential from one time point to the next. However, there might be additional variation from this global In this chapter, we laid the framework to model how data that lie on a simplex change over time. reassurance that our two samplers are largely converging to the same estimates of \( \eta \) and \( v \).

### 2.6 Conclusion

In this chapter, we laid the framework to model how data that lie on a simplex change over time. We first introduced the affine transformation, which is a modified affine combination between two points. Such a transformation is useful for explicating a global trend in how all data points change from one time point to the next. However, there might be additional variation from this global trend. To model this "noise", we introduced the additive Dirichlet error model, the exponential rotational, and the rotational geodesic error model. We also discussed how to perform inference on these models. Out of the three models, the rotational geodesic error model seemed most promising, particularly after our simulation studies. We found that we were able to recover the general affine transformation parameter when the rotational geodesic error model was combined with the general affine transformation. Using these insights from the simulation studies, we analyzed the Los Angeles County income proportion data set with these models. The trends revealed agreed with larger macroeconomic ones. The model additionally illustrated how strong they were, highlighting the utility of interpretability.

<table>
<thead>
<tr>
<th>Year</th>
<th>( \eta )</th>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>( v_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990 - 1991</td>
<td>-0.079</td>
<td>0.002</td>
<td>0.251</td>
<td>0.748</td>
</tr>
<tr>
<td>1991 - 1992</td>
<td>-0.073</td>
<td>0.002</td>
<td>0.571</td>
<td>0.427</td>
</tr>
<tr>
<td>1992 - 1993</td>
<td>0.250</td>
<td>0.880</td>
<td>0.077</td>
<td>0.043</td>
</tr>
<tr>
<td>1993 - 1994</td>
<td>-0.057</td>
<td>0.012</td>
<td>0.332</td>
<td>0.656</td>
</tr>
<tr>
<td>1994 - 1995</td>
<td>-0.0001</td>
<td>0.380</td>
<td>0.321</td>
<td>0.299</td>
</tr>
<tr>
<td>1995 - 1996</td>
<td>0.112</td>
<td>0.769</td>
<td>0.100</td>
<td>0.131</td>
</tr>
<tr>
<td>1996 - 1997</td>
<td>0.089</td>
<td>0.615</td>
<td>0.166</td>
<td>0.218</td>
</tr>
<tr>
<td>1997 - 1998</td>
<td>0.074</td>
<td>0.321</td>
<td>0.432</td>
<td>0.247</td>
</tr>
<tr>
<td>1998 - 1999</td>
<td>0.028</td>
<td>0.209</td>
<td>0.690</td>
<td>0.101</td>
</tr>
<tr>
<td>1999 - 2000</td>
<td>0.047</td>
<td>0.048</td>
<td>0.818</td>
<td>0.133</td>
</tr>
<tr>
<td>2000 - 2001</td>
<td>0.070</td>
<td>0.330</td>
<td>0.247</td>
<td>0.423</td>
</tr>
<tr>
<td>2001 - 2002</td>
<td>0.062</td>
<td>0.082</td>
<td>0.795</td>
<td>0.122</td>
</tr>
<tr>
<td>2002 - 2003</td>
<td>0.048</td>
<td>0.766</td>
<td>0.106</td>
<td></td>
</tr>
<tr>
<td>2003 - 2004</td>
<td>0.191</td>
<td>0.834</td>
<td>0.159</td>
<td></td>
</tr>
<tr>
<td>2004 - 2005</td>
<td>0.330</td>
<td>0.823</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td>2005 - 2006</td>
<td>0.364</td>
<td>0.705</td>
<td>0.223</td>
<td></td>
</tr>
<tr>
<td>2006 - 2007</td>
<td>0.154</td>
<td>0.543</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>2007 - 2008</td>
<td>0.554</td>
<td>0.198</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>2008 - 2009</td>
<td>-0.137</td>
<td>0.473</td>
<td>0.523</td>
<td></td>
</tr>
<tr>
<td>2009 - 2010</td>
<td>-0.011</td>
<td>0.065</td>
<td>0.059</td>
<td>0.876</td>
</tr>
</tbody>
</table>

Table 2.1: Mean posterior estimates for \( \eta \) and \( v \), starting at \( (1/3, 1/3, 1/3) \). Results from sampling all indices of \( v \) before \( \eta \) are on the left and Results from sampling an index of \( v \) and then \( \eta \) are on the right.
CHAPTER 3

Simplex Random Direction

3.1 Introduction

The previous chapter assumed that the random direction associated with the noise model is uniform. However, as demonstrated by the Los Angeles County income proportion data set, this might not be appropriate. Indeed, this modeling task presents potential non-trivial difficulties. As we will demonstrate later, this shift appears dependent on the current income proportions and not on the neighborhood’s physical location. Then, Figure 3.1 shows that we cannot assume a uniform random direction in this change because there may be clearly preferred directions. We need a way to assign probability to these random directions. In order to start thinking about how to accomplish this, we first examine the 2D simplex case, which shall be the main focus of our chapter. There, the simplex is essentially a two dimensional surface for our points to move in. It suffices to assume that these random directions are random angles and lie in the interval \([0, 2\pi]\). One naive approach to assign probability to such an interval is by assigning probability to some random variable \(z \in \mathbb{R}\) and using the inverse logit function to transform \(z\) to that interval. Such an approach would be problematic because the endpoints of the interval and values near the end points are mapped near their respective \(\pm \infty\) and are far apart. Meanwhile, the end points for the random angle’s interval, 0 and 2\(\pi\), denote the same direction and should not be so far apart. The issue here, mathematically, is that the inverse logit function that maps \(\mathbb{R}\) to the angles \([0, 2\pi]\), while continuous and one-to-one, does not possess a continuous inverse.

Existing work Several useful techniques for modeling (random) angles and directions come from directional statistics [45]. In particular, we can think of the random angle as circular data because these angles can be mapped to a point on the unit circle and circular data are observations that lie on a unit circle. There are several basic distributions on circular data that we can utilize for our purposes: the von Mises distribution and the projected normal distribution in two dimensions. The former can be loosely viewed as the circular version of the normal distribution [40]. Meanwhile, the latter is a bivariate Gaussian distribution that is transformed to a distribution on the circle using the
Figure 3.1: Random direction plots for different years. The \(x-y\) coordinates represent the income proportion in the first two categories. If the plot is three dimensional, the \(z\) coordinate is the random direction divided by \(2\pi\) to lie in \([0, 1)\). The two leftmost plots show all random directions for the years listed below. The middle plots show the movement for a subset of 20-30 "locations" and the corresponding extracted directions. The two rightmost plots display a histogram of the random directions associated with locations that are within 0.05 \(L_2\)-distance of \((0.39, 0.56, 0.05)\) for 2003-2004 and of \((0.85, 0.10, 0.05)\) for 2008-2009.

usual polar coordinate transform and integrating out the radius [45]. There is another distribution that we might use. As seen in Figure 3.1, it is possible that nearby locations move in mostly similar directions. In other words, the random directions, depending on where the start points are, appear to be spatially correlated. Because the change in the pattern of random direction is also smooth as we pass across the simplex in this figure, an appropriate model might be the Gaussian process [59]. Even though we cannot use the inverse logit function to transform a Gaussian process because of what we discussed earlier about angles, we can use the circular version of the Gaussian process, i.e. the projected Gaussian process [70].

**Our approach** To model the random direction of movements for simplex-valued data, we will leverage and expand upon the building blocks advocated by [45, 59] and the techniques of [70, 71]. As the first step of our approach, we assume that the observed random directions are distributed according to a von Mises distribution. We then correlate the von Mises distributions’ means with a projected Gaussian process. In other words, this is a circular version of the Gaussian process with Gaussian white noise. This modeling choice seems appropriate because there appears to be an
unimodal empirical distribution for the random directions associated with the nearby locations for any location in the simplex. For example, Figure 3.1d shows such a distribution for the location in the simplex, \((0.39, 0.56, 0.05)\). Such a choice also makes sense because the mean of the von Mises distribution can be thought of as a vector on the unit circle, which is what a projected Gaussian process outputs. In addition, it allows both the prior and likelihood to recognize the geometry of angles. Not only can this basic model harness the power of Gaussian processes to spatially correlate random directions of nearby locations and handle noisy directions, but also it does so in an interpretable way. After fitting our model, we can make a posterior prediction of the mean preferred direction for any location.

Extensions to the basic model described above are required to account for "heterogenous" patterns for random directions, which are evidently illustrated in Figure 3.1e. Because the basic approach proves to be inadequate, the modeling extensions form a substantial portion of this chapter. To capture such heterogeneity, we shall extend the basic model via several mixture modeling techniques. In particular, the observation is distributed according to one of \(K \geq 1\) von Mises distributions with some probability. Each of the \(K\) distributions has a projected Gaussian process to correlate its mean. This simple extension still retains the advantages mentioned previously because each component now represents a mean preferred direction. We also discuss another modeling alternative. While there still are \(K\) von Mises distributions, we extend the inverse logit to transform \(K-1\) Gaussian processes to the mixing probability for each component.

We have applied and evaluated these models to a simulated data set and the aforementioned data set of income proportions. For model selection, we compare the fitted models’ posterior predictive probabilities calculated on data withheld from each year. The chosen models enable us to discover several patterns of interest on the year-to-year income proportion movements. There appear to be four phases based on the patterns we discover, and interestingly enough, these phases correspond to the economic cycles observed during that time period. The first two phases charts the end of one cycle, the third phase follows the dot com bubble and the recession after the bubble, and the last phase represents the subsequent housing bubble and the housing market crash. Because we can also assign meaning to the random directions with respect to the income categories, interpreting the results from the models gives us further information about the year to year change during these phases.

These modeling techniques can also be extended to random directions in higher dimensions. In addition, it might be applied to data that lie on manifolds other than the simplex because any change in the data can be represented as a random movement, which can be decomposed accordingly. However, in this chapter, we shall restrict our attention to two dimensional simplex data in order to focus on the heterogeneous modeling and inference in the circular representation, which is already quite intricate. The rest of the chapter is organized as follows. In Section 3.2, we describe how the
polar coordinate transformation can be used to transform distributions in $\mathbb{R}^2$ to distributions on the circle, such as the von Mises distribution or the projected normal distribution in two dimensions. We also discuss some properties of these transformed distributions and extend the projected bivariate normal distribution to the projected Gaussian process. While it is not a focus of this chapter, we then introduce the higher dimension version of these distributions. Next, we more formally introduce our models in the Section 3.3 and describe a few properties of these models using ideas from directional statistics. We then briefly examine how to fit these models in Section 3.4. We fit them to simulated and real data in Section 3.5 and discuss these results. Section 3.6 discusses possible extensions.

3.2 Polar coordinate transformations for circular data

In this section, we examine how the polar coordinate transformation can be used to transform distributions in Euclidean domains, such as $\mathbb{R}^2$, to distributions on non-Euclidean domains, such as the unit circle. Because our chapter is focused on random directions in $\Delta^2$, we begin by discussing tools from circular statistics and then introduce the higher dimension analogues. In circular statistics, the data are observations that lie on an unit circle [45], so they can be represented by their angles, which will be denoted by random variable $Y \in [0, 2\pi)$.

Given any $(z_1, z_2) \in \mathbb{R}^2 \setminus \{0\}$, there is a unique polar representation in terms of the corresponding radius and angles, namely $(r, y) \in \mathbb{R}^+ \times [0, 2\pi)$, where $z_1 = r \cos y$ and $z_2 = r \sin y$. In fact, this representation establishes a bijection between $(r, y) \in \mathbb{R}^+ \times [0, 2\pi)$ and $(z_1, z_2) \in \mathbb{R}^2 \setminus \{0\}$. This suggests a natural recipe for assigning probability for (random) angles $y \in [0, 2\pi)$: this can be done by endowing a distribution for $(z_1, z_2)$, which induces a distribution for the pair $(r, y)$. We obtain a valid distribution on angles $[0, 2\pi)$ by either marginalizing out $r$ or conditioning on a particular value of $r$.

The aforementioned bijection between $(z_1, z_2)$ and $(r, y)$ induces a function that maps $(z_1, z_2) \in \mathbb{R}^2 \setminus \{0\}$ to $y \in [0, 2\pi)$. This function is denoted by $y = \arctan^*(z_1, z_2)$, where the notation $\arctan^*$ is used to indicate a modified version of the standard arctan function. Recall that the standard arctan function maps the real line $\mathbb{R}$ to the open interval $(-\pi/2, \pi/2)$. We may extend this continuous function to the closed domain $\mathbb{R} := [-\infty, +\infty]$ and the closed range of angles $[0, 2\pi)$. The function $\arctan^*$ satisfies the following (which is commonly used as a definition): for $(z_1, z_2) \in \mathbb{R}^2 \setminus \{0\}$,

$$\arctan^*(z_1, z_2) = \begin{cases} 
\arctan\left(\frac{z_2}{z_1}\right) & z_1 \geq 0, z_2 \geq 0 \\
\arctan\left(\frac{z_2}{z_1}\right) + 2\pi & z_1 \geq 0, z_2 < 0 \\
\arctan\left(\frac{z_2}{z_1}\right) + \pi & z_1 < 0.
\end{cases} \quad (3.1)
$$

Using these representations, we can derive the von Mises and projected normal distributions,
which will be useful building blocks in our subsequent modeling.

### 3.2.1 von Mises distributions

A natural choice for constructing a probability distribution of \((z_1, z_2) \in \mathbb{R}^2\) is the bivariate Gaussian distribution with mean, \((\cos(\alpha), \sin(\alpha))^T\), and variance, \(\rho^{-1} \mathbb{I}_{2 \times 2}\), for some \(\alpha \in [0, 2\pi)\) and \(\rho \in \mathbb{R}^+\). The precise meaning of parameters \(\alpha\) and \(\rho\) will be clear shortly. Applying change of variables, we obtain that for \(r > 0\),

\[
p(r, y \mid (\cos(\alpha), \sin(\alpha))^T, \rho^{-1} \mathbb{I}_{2 \times 2}) = r \frac{\rho}{2\pi} \exp \left( -\frac{\rho}{2} \left( (r \cos(y) - \cos(\alpha))^2 + (r \sin(y) - \sin(\alpha))^2 \right) \right)
\]

\[
= r \frac{\rho}{2\pi} \exp \left( -\frac{\rho}{2} (r^2 - 2r \cos(y - \alpha) + 1) \right)
\]

\[
p(r, y \mid (\cos(\alpha), \sin(\alpha))^T, \rho^{-1} \mathbb{I}_{2 \times 2}) = r \frac{\rho}{2\pi} \exp \left( -\frac{\rho}{2} (r^2 - 2r \cos(y - \alpha) + 1) \right).
\]

By conditioning on \(r = 1\), we arrive at the von Mises distribution on the interval \([0, 2\pi)\) [45]:

\[
p(y \mid r = 1, \alpha, \rho) \propto \exp(\rho \cos(y - \alpha)).
\]

In its original form, the von Mises distribution is defined by two parameters: a mean angle, \(\alpha \in [0, 2\pi)\), and a concentration parameter, \(\rho \in \mathbb{R}^+\). Given these parameters, the von Mises distribution has a density function of the following form:

\[
f_V(y \mid \alpha, \rho) := \frac{e^{\rho \cos(y - \alpha)}}{2\pi I_0(\rho)}.
\]

(3.2)

Appearing in the normalizing constant, \(I_0(\rho)\) is the modified Bessel function of the first kind and of order 0. In general, the modified Bessel function of the first kind and of order \(n\) on the interval of \([0, 2\pi)\) is defined to be

\[
I_n(\rho) := \frac{1}{2\pi} \int_{0}^{2\pi} \cos(ny) e^{\rho \cos(y)} dy.
\]

(3.3)

Since the domain of the angles, the unit circle, is non-Euclidean, one has to be careful when speaking of notions such as mean and variance. Here are several properties we wish to highlight. First, the distribution is unimodal and symmetric around its mean, \(\alpha\), if \(\rho > 0\). This makes sense because as discussed earlier, the density function is proportional and restricted to a bivariate Gaussian centered at a point on the unit circle. If \(\rho = 0\), then the von Mises distribution becomes the uniform distribution on an interval of length \(2\pi\). Next, the mean angle, \(\alpha\), is different from the
typical definition because for the von Mises distribution with $\rho > 0$, it is the $\alpha \in [0, 2\pi)$ such that the following quantity

$$E(e^{iY}) := \int e^{iy} f_Y(y \mid \alpha, \rho) \, dy,$$

(3.4)

where $i$ denotes the complex number, satisfies the following identity

$$E(e^{iY}) = \frac{I_1(\rho)}{I_0(\rho)} e^{i\alpha}.$$  

(3.5)

Intuitively, the circular mean is a weighted average of points on the circle and the angles that correspond to them instead of the values of the angles. Indeed, if $f_C(y)$ is some distribution on the circle, the circular mean in general is the $\alpha \in [0, 2\pi)$ such that

$$E(e^{iY}) := \int e^{iy} f_C(y) \, dy$$

(3.6)

satisfies the following equation for some $r \in \mathbb{R}^+$,

$$E(e^{iY}) = re^{i\alpha}.$$  

(3.7)

For convenience, we may refer to the circular mean as the mean or average later on. It should be obvious from context if we refer to the typical or circular mean.

Finally, despite the circular variance still using the mean direction $\alpha$, the circular variance is different as well. It is defined on the interval $[0, 2\pi)$ to be

$$\text{Var} (Y) := 1 - \int \cos(y - \alpha) f_C(y) \, dy.$$  

(3.8)

We may also write $\text{Var} (Y) = 1 - E(\cos(Y - \alpha))$ [45], where the left hand side refers to the circular variance, and the right hand side makes use of the usual expectation (of a scalar random variable). It should be obvious from context which type of variance we are discussing. Due to how circular variance is defined, it takes values between 0 and 1. At a high level, if $Y$ is tightly concentrated around the mean direction $\alpha$, the variance will be close to zero. Conversely, if $Y$ is dispersed around the circle, $E(\cos(Y - \alpha))$ will be 0 so the variance will be 1. Because the variance indicates the spread of the data, $E(\cos(Y - \alpha))$ can be seen as a measure of concentration for the distribution $f_C$. For a von Mises distribution, the circular variance is

$$\text{Var} (Y) = 1 - \frac{I_1(\rho)}{I_0(\rho)}.$$  

(3.9)
3.2.2 Projected normal distribution

The von Mises distribution is not the only distribution that we can derive from transforming a bivariate Gaussian distribution to a distribution on random angles. Let us consider a bivariate Gaussian with any mean, \( \mu \in \mathbb{R}^2 \), covariance matrix, \( \Sigma \). Then, if we apply the polar coordinate transformation to this distribution, we have that

\[
p(r, y \mid \mu, \Sigma) = \frac{r}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}((r \cos(y), r \sin(y))^T - \mu)^T \Sigma^{-1} (r \cos(y), r \sin(y))^T - \mu\right).
\]

The projected normal distribution in two dimensions immediately follows from this because it is the marginal distribution of \( Y \), i.e.

\[
f_N(y \mid \mu, \Sigma) = \int_{r>0} p(r, y \mid \mu, \Sigma) dr.
\]  

Even though this integral is tractable [45], we focus on the case \( \Sigma = I_{2 \times 2} \) because its properties are well-understood. Given \( \mu \), let \( \mu_0 \in \mathbb{R}^+ \) and \( \alpha \in [0, 2\pi) \) be the polar coordinate transform of \( \mu \). As shown by Wang and Gelfand [70], the distribution in this scenario has the following closed-form expression:

\[
f_N(y \mid \mu, I_{2 \times 2}) = \int_{r>0} \frac{r}{2\pi} \exp\left(-\frac{1}{2}((r \cos(y) - \mu_0 \cos(\alpha))^2 + (r \sin(y) - \mu_0 \sin(\alpha))^2)\right) dr
\]

\[
= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu_0^2 \sin^2(y - \alpha)\right)
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{r>0} r \exp\left(-\frac{1}{2}(r - \mu_0 \cos(y - \alpha))^2\right) dr
\]

\[
= \phi(\mu_0 \sin(y - \alpha))(\phi(\mu_0 \cos(y - \alpha)) + \mu_0 \cos(y - \alpha)\Phi(\mu_0 \cos(y - \alpha))).
\]  

Here, \( \phi(x) \) and \( \Phi(x) \) denote the standard normal PDF and CDF. If we also then set \( \mu = 0 \), the projected normal distribution becomes uniform. For \( \mu \neq 0 \), the projected normal distribution with the identity matrix is symmetric and unimodal around \( \alpha \) like the von Mises distribution. It is not surprising that based on this fact, the circular mean is \( \alpha \) [70]. If \( \alpha \) is undefined due to \( \mu = 0 \), then the mean angle is \( \pi \) because the projected normal distribution is a uniform distribution. Meanwhile, we can use Kendall’s technique to compute the circular variance [37]. Despite Kendall only considering the case in which \( \mu_1 \in \mathbb{R}^+ \) and \( \mu_2 = 0 \), it is straightforward to extend his result and show that the circular variance is

\[
1 - \frac{1}{2} \sqrt{2\pi} \beta \exp(-\beta)(I_0(\beta) + I_1(\beta)),
\]

where \( \beta = \frac{\mu_0^2}{4} \) and \( I_0(\beta) \) and \( I_1(\beta) \) are the modified Bessel functions defined in (3.3). Again, like the von Mises distribution, the circular variance is independent of the mean angle. Then, we can use the asymptotics of the Bessel function
to show that the circular variance is 1 when $\beta \to 0$ and 0 when $\beta \to \infty$ [72]. As discussed before, the former occurs when $\mu_0 \to 0$ and the distribution becomes uniform. The latter happens when $\mu_0 \to \infty$ and the probability (density) of other angles tend to 0.

We cannot use these results to derive closed-form expressions for the circular mean and variance in the general case because the projected normal distribution in that case has a more complicated closed-form expression. Fortunately, as we will demonstrate later in this chapter, these results do allow us to compute the circular mean and variance of the models that we will introduce.

### 3.2.3 Projected Gaussian process

It is of interest to define a stochastic process of random directions indexed in a general domain $\Omega$. This is possible by exploiting the polar coordinate transformation as above, whereas the bivariate normal distribution for $(z_1, z_2)$ can be extended into a stochastic process defined on $\Omega$; a common choice is to use the Gaussian process. This idea was studied by [71]. We will present a simpler version of this idea.

Gaussian process is a powerful modeling tool for spatio-temporal data [18, 9, 59]. Given observations $z_1, z_2, \ldots, z_N$ indexed by the corresponding locations $x_1, x_2, \ldots, x_N \in \Omega$, we assume that these observations are realizations of a stochastic process $\{Z(x)|x \in \Omega\}$, i.e., $z_\ell = Z(x_\ell)$ for $\ell = 1, 2, \ldots, N$. To account for the spatial dependence of these observations, one may assume that $Z$ is a Gaussian process, which is parameterized by a mean function $m$ and a covariance function $K$ on $\Omega$. Abusing notation, we let $m$ denote the mean function applied to every location such that $m_\ell = m(x_\ell)$. If $\Sigma$ is an $N \times N$ matrix such that $\Sigma_{\ell, \ell'} = K(x_\ell, x_{\ell'})$ for $\ell, \ell' \in 1, 2, \ldots, N$, we will denote this as $Z \sim GP(m, \Sigma)$. We write it this way because $p(Z | m, \Sigma) = \mathcal{N}(Z | m, \Sigma)$.

We describe how to convert the Gaussian process to the projected Gaussian process. Suppose that $Z_1 \sim GP(m_1, \Sigma_1)$ and $Z_2 \sim GP(m_2, \Sigma_2)$ with $Z_1, Z_2 \in \mathbb{R}^N$. For identifiability of the angle-valued stochastic process, we might assume that

\[
\begin{pmatrix}
\Sigma_1 & 0 \\
0 & \Sigma_2
\end{pmatrix}
\] is equal to

\[
\begin{pmatrix}
1 & 0 \\
0 & \sigma^2
\end{pmatrix}
\otimes \Sigma
\]

for some fixed $\sigma^2 > 0$ and covariance matrix, $\Sigma$. While this may appear restrictive, it is similar in spirit to what Wang and Gelfand proposed to identify the projected normal distribution in two dimensions [70]. Then, let $\mathcal{A} \in [0, 2\pi)^N$ denote the angle-valued stochastic process observed at the $N$ index points so that $\mathcal{A}_\ell = \mathcal{A}(x_\ell)$ for $\ell = 1, 2, \ldots, N$ and $r \in (\mathbb{R}^+)^N$ be a latent variable. We apply the polar coordinate transformation element-wise such that $z_{1, \ell} = r_\ell \cos(\mathcal{A}_\ell)$ and $z_{2, \ell} = r_\ell \sin(\mathcal{A}_\ell)$ for $\ell = 1, 2, \ldots, N$. As a shorthand, define $\cos(\mathcal{A})$ and $\sin(\mathcal{A})$ to be the vectors such that $\cos(\mathcal{A})_\ell = \cos(\mathcal{A}_\ell)$ and $\sin(\mathcal{A})_\ell = \sin(\mathcal{A}_\ell)$ for $\ell = 1, 2, \ldots, N$. Since the Jacobian of
this transformation is $\prod_\ell r_\ell$, we have that
\[
p(\mathcal{A}, r \mid m_1, m_2, \Sigma_1, \Sigma_2) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_1|}} \exp \left( -\frac{1}{2} (r \cos(\mathcal{A}) - m_1)^T \Sigma_1^{-1} (r \cos(\mathcal{A}) - m_1) \right) \]
\[
\frac{1}{\sqrt{(2\pi)^N |\Sigma_2|}} \exp \left( -\frac{1}{2} (r \sin(\mathcal{A}) - m_2)^T \Sigma_2^{-1} (r \sin(\mathcal{A}) - m_2) \right) \left( \prod_\ell r_\ell \right)
\]
(3.12)
The projected Gaussian process is the marginal distribution of $\mathcal{A}$:
\[
p(\mathcal{A} \mid m_1, m_2, \Sigma_1, \Sigma_2) = \int_{r > 0} p(\mathcal{A}, r \mid m_1, m_2, \Sigma_1, \Sigma_2) dr.
\]
(3.13)
While this integral is intractable, we can still calculate its properties marginally for $A_\ell$ to understand this process. Because for $\ell = 1, 2, \ldots, N$, $\mathcal{A}_\ell$ is marginally distributed according to $f_N \left( (m_{1,\ell}; m_{2,\ell})^T, \begin{pmatrix} (\Sigma_1)_{\ell,\ell} & 0 \\ 0 & (\Sigma_2)_{\ell,\ell} \end{pmatrix} \right)$. If $(\Sigma_1)_{\ell,\ell}, (\Sigma_2)_{\ell,\ell} = 1$, we can use the previous section’s discussion to do so. Otherwise, we will explain the computation later.

### 3.2.4 Higher dimension analogues

An alternative way to describe these distributions is to view them as distributions on vectors "projected" onto the unit sphere. In other words, given $z \in \mathbb{R}^2$, the von Mises distribution and the projected normal distribution in two dimensions are distributions on $z/\|z\|_2$. From this perspective, because $r = \|z\|_2$ in the polar coordinate transform, it further explains why the von Mises distribution is proportional to a particular bivariate Gaussian conditioned on $r = 1$. Meanwhile, the projected normal distribution arises by computing the probability of $z/\|z\|_2$ if $z$ is distributed according to a bivariate Gaussian with mean $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^2 \times \mathbb{R}^2$. In other words, we have that
\[
f_N \left( \frac{z}{\|z\|_2} \mid \mu, \Sigma \right) = \int_{z' \mid \|z'\|_2 = \|z\|_2} p(z' \mid \mu, \Sigma) dz'.
\]
(3.14)
This calculation is easier to do with polar coordinates.

This perspective also suggests how to generalize the von Mises distribution and the projected normal distribution for higher dimensions. First, for the von Mises distribution, suppose that $z \in \mathbb{R}^D$, $D > 2$. Let $z$ be distributed according to a Gaussian distribution of dimension $d$ with mean $\mu$ and covariance matrix, $\rho^{-1} I_D$. Here, $\mu \in \mathbb{R}^D$ such that $\|\mu\|_2 = 1$ and $\rho \in \mathbb{R}^+$. Then,
\[
p(z \mid \mu, \rho^{-1} I_D) = \left( \frac{\rho}{2\pi} \right)^{\frac{d}{2}} \exp \left( -\frac{\rho}{2} (z - \mu)^2 \right)
\]
\[ p(z \mid \|z\|_2 = 1, \mu, \rho) \propto \exp(\rho \mu^T z). \]

The normalization constant for this distribution is \( (\rho/2\pi)^D \rho^D \pi^{D-1} \rho I_{D-1}(\rho) \) where \( I_{D-1}(\rho) \) is the modified Bessel function defined in (3.3) [45]. Note that this is identical to the von Mises distribution in two dimensions because if \( \mu = (\cos(\alpha), \sin(\alpha)) \) and \( z = (\cos(y), \sin(y)) \) for \( y, \alpha \in [0, 2\pi) \),

\[
\mu^T z = (\cos(\alpha), \sin(\alpha))^T (\cos(y), \sin(y))
= \cos(\alpha) \cos(y) + \sin(\alpha) \sin(y)
= \cos(y - \alpha).
\]

The projected normal distribution in higher dimension can be derived in a similar manner from its lower dimensional analogue. The projected normal distribution is given by computing the probability of \( z \) if \( z \) is now distributed according to a Gaussian of dimension \( D \) with mean, \( \mu \in \mathbb{R}^D \), and covariance matrix, \( \Sigma \in \mathbb{R}^{D \times D} \). The calculation is the same as (3.14) except with \( \mu \) and \( \Sigma \) of the appropriate dimension. While this distribution might not be identifiable [70], \( \Sigma \) can be set in certain ways to simplify fitting the distribution to data [31].

This then provides a template for defining the projected Gaussian process in higher dimensions. Suppose that for all \( d = 1, 2, \ldots, D \), \( Z_d \sim GP(\mu_d, \Sigma_d) \). Again, \( \mu_d \in \mathbb{R}^N \) and \( \Sigma_d \in \mathbb{R}^{N \times N} \) for any \( d \). Set \( C \in \mathbb{R}^N \times \mathbb{R}^D \), \( \|C_n\| = 1 \) to be a stochastic process on the unit circle. We then element-wise project each to the unit circle such that \( C_{n,d} = \frac{Z_{d,n}}{\sqrt{\sum_{d'} Z_{d',n}^2}} \). If we abuse notation and denote this element-wise projection as \( C(Z_1, Z_2, \ldots, Z_D) \), then the probability is the following:

\[
p(C \mid \mu_1, \mu_2, \ldots, \mu_D, \Sigma_1, \Sigma_2, \ldots, \Sigma_D) = \int_{C(Z_1, Z_2, \ldots, Z_D) = C} \prod_d p(Z_d \mid \mu_d, \Sigma_d) dZ_d.
\]

### 3.3 Modeling random directions

#### 3.3.1 Model description

The purpose of our models is to elucidate the patterns observed in random directions. We first introduce the notation that we intend to use.

We are given a location \( x_\ell \in \Delta^2 \) and an observation \( y_\ell \in [0, 2\pi) \) for that location for \( \ell = 1, 2, \ldots, N \). We use \( \ell \) as our index because of the spatial aspects of the data and model. To
Table 3.1: A table showing the relationship between the models discussed in Section 3.3.1 and the supplementary material. The models introduced in this chapter are in bold.

<table>
<thead>
<tr>
<th>Observed Pattern</th>
<th>Homogeneous</th>
<th>Heterogeneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ind.</td>
<td>(iV) (B.1)</td>
<td>(iVM) (B.2)</td>
</tr>
<tr>
<td>Var.</td>
<td>(SvM) (B.3)</td>
<td>(SvM-c) (3.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(SvM-p) (3.16), (3.17)</td>
</tr>
</tbody>
</table>

compare two observations, we will use \(\ell'\) to indicate a location and observation different from the location and observation corresponding to \(\ell\). If there are \(K\) von Mises distributions indexed by \(k, k \in 1, 2, \ldots, K\), each observation, \(y_\ell\), is assumed to be distributed according to one of the von Mises distributions with mean parameter, \(m_{k,\ell} \in [0, 2\pi]\), and a concentration parameter, \(\rho_{k,\ell} \in \mathbb{R}^+\). We will use \(\zeta_\ell\) to indicate which von Mises observation is associated with. If there is one von Mises distribution, we will drop \(k\) from the parameters’ index and will not use \(\zeta_\ell\). We will remove \(\ell\) from the index if the parameters are the same regardless of the location or if we wish to refer to the entire vector. While this overloading is unfortunate, any vector will be displayed in bold. Even though context should make it clear which definition we are referring to, we will specify the space the variable lies in.

We also need a set of notation for Gaussian processes because all the models that we will introduce use it. We let \(z_k \in \mathbb{R}^N\) be a draw from a Gaussian process with mean \(\mu_k \in \mathbb{R}^N\) and covariance matrix \(\Sigma_k\). For models that need two Gaussian processes for each component, we will further index these vectors as \(z_{k,1}\) and \(z_{k,2}\) and their parameters as \(\mu_{k,1}, \mu_{k,2}, \Sigma_{k,1},\) and \(\Sigma_{k,2}\). The specific value corresponding to observation \(y_\ell\) will be denoted by \(z_{k,\ell}\) or \(z_{k,1,\ell}\) and \(z_{k,2,\ell}\). Again, we may drop \(k\) if there is only one Gaussian process. In both cases, our choice of indices is to make the link between the specific Gaussian process and the particular von Mises distribution and observation more explicit.

**Heterogenous Spatial Random Direction Model** There are two ways to incorporate spatial information for heterogeneous spatial random direction patterns. The first way is to integrate spatial information in the means of the von Mises distributions. We call this model the *spatially varying von Mises component mixture* model or \(SvM-c\). More specifically,

\[
z_{k,1} \sim \text{GP}(\cdot \mid \mu_{k,1}, \Sigma_k), \quad k = 1, 2, \ldots, K
\]
\[ z_{k,2} \sim \text{GP}(\cdot \mid \mu_{k,2}, \Sigma_k), \quad k = 1, 2, \ldots, K \]
\[ m_k = \arctan^*(z_{k,1}, z_{k,2}), \quad k = 1, 2, \ldots, K \]
\[ \varphi_k \overset{iid}{\sim} \mathcal{N}(\cdot \mid \nu_k, \varsigma^2), \quad k = 1, 2, \ldots, K \]
\[ \rho_k = \exp\{\varphi_k\}, \quad k = 1, 2, \ldots, K \]

\[ \zeta_\ell \mid \lambda_1, \lambda_2, \ldots, \lambda_K \overset{iid}{\sim} \text{Cat}(\cdot \mid \lambda_1, \lambda_2, \ldots, \lambda_K) \]
\[ y_\ell \mid \zeta_\ell = k, m_{k,\ell}, \rho_{k,\ell} \sim f_Y(\cdot \mid m_{k,\ell}, \rho_{k,\ell}), \quad \ell = 1, 2, \ldots, N. \]

According to this model specification, each observation may be distributed by one of \( K \) von Mises distributions with probability \( \lambda_k \) regardless of its location. Each distribution’s mean parameters, \( m_k \in [0, 2\pi)^N \), are transformed from two draws, \( z_{k,1}, z_{k,2} \in \mathbb{R}^N \), from one of \( K \) Gaussian process with potentially its own mean \( \mu_k \in \mathbb{R}^N \) and covariance matrix \( \Sigma_k \). This transformation is accomplished using the \( \arctan^* \) function element-wise. A distribution’s concentration parameters, \( \rho_k \), are again random variables, \( \varphi_k \), that have been exponentiated. These random variables are distributed according to a hierarchical normal distribution. At a lower level, they are distributed according to a normal distribution with the same standard deviation, \( \varsigma \), but with different hierarchical means, \( \nu_k \). These hierarchical means, \( \nu_k \), are given the same hyperprior, \( \mathcal{N}(0, \tau) \).

In other words, for the concentration parameter in \( \text{SvM-c} \), we use the following hierarchical prior:

\[ \nu_k \sim \mathcal{N}(\cdot \mid 0, \tau^2), \]
\[ \varphi_k \overset{iid}{\sim} \mathcal{N}(\cdot \mid \nu_k, \varsigma^2). \]

We use the hierarchical prior for the concentration parameter because it is a compromise between assigning an individual and a global concentration parameter. Using a global parameter \( \rho \) might affect the estimates of the mean if the variances differ significantly because the model cannot adjust the concentration parameter. Conversely, assigning an individual parameter makes the model too flexible. This might negatively affect the model’s ability to spatially correlate the observations. This concern also leads us to set the standard deviation for the lower term, \( \varsigma \), to a small value instead of sampling for it. Even with a tight prior on \( \varsigma \), the variance of the lower terms will be greater if we sample for the standard deviation. In addition, we do not use another Gaussian process to model the variance parameter \( \varphi_\ell \) for computation reasons and for fears of making the model too rich. Still, a normal distribution is useful because it will allow us to separately sample the hierarchical mean, \( \nu_k \), from the lower term, \( \varphi_{k,\ell} \).

\( \text{SvM-c} \) included spatial information through the components. An alternative approach to meld spatial information into our models is through the mixing probability. In this approach, we assume
that observations nearby are likely to belong to the same von Mises distribution instead of being distributed according to a von Mises distribution with similar means. This alternative model is still interpretable because the results describe the preferred von Mises distributions of certain regions.

More explicitly, we call this model the *spatially varying mixing probability for von Mises distributions* model. Keeping with our previous convention, we will shorten it to *SvM-* $p$. For this model, $y_\ell$ is assumed to be distributed according to one of $K$ von Mises distribution with mean parameter, $m_k \in [0, 2\pi)$, and concentration parameter, $\rho_k \in \mathbb{R}^+$. Both parameters are assumed to be the same for a particular von Mises distribution regardless of the observations’ location. The mean parameters across all von Mises distributions are consequently given a prior of $\text{Unif}(0, 2\pi)$ whereas the concentration parameters are given a prior of $\Gamma(1, 1)$. If there is more prior knowledge about either parameter or if we want to force identifiability among the parameters, a von Mises distribution and a stronger Gamma distribution could instead be used for the mean parameter and concentration parameter respectively. Meanwhile, because the mixing probabilities sum up to 1, we only need to use the generalized inverse logit function, $\Psi^{-1}()$, to element-wise transform draws from $K - 1$ Gaussian processes with mean $\mu_k$ and covariance matrix $\Sigma_k$ to the mixing probabilities. We define the generalized inverse logit function, $\Psi^{-1}(z_1, z_2, \ldots, z_{k-1})$, as a mapping from $\mathbb{R}^{K-1}$ to $\Delta^K$ in the following manner:

\[
\Psi^{-1}(z_1, z_2, \ldots, z_{k-1})_k = \frac{\exp(z_{k,\ell})}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})}, \quad k = 1, 2, \ldots, K - 1
\]

\[
P_{si}^{-1}(z_1, z_2, \ldots, z_{k-1})_K = \frac{1}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})}.
\]

We will still use $k$ as our index for the Gaussian process draws to make clear which draw links to which von Mises distribution. In short, *SvM-* $p$ can be described as following.

\[
m_k \sim \text{Unif}(. \mid 0, 2\pi), \quad k = 1, 2, \ldots, K
\]

\[
\rho_k \sim \Gamma(. \mid 1, 1), \quad k = 1, 2, \ldots, K
\]

\[
z_k \sim \text{GP}(. \mid \mu_k, \Sigma_k), \quad k = 1, 2, \ldots, K - 1
\]

\[
\lambda_1, \lambda_2, \ldots, \lambda_K = \Psi^{-1}(z_1, z_2, \ldots, z_{k-1})
\]

\[
\zeta_\ell \mid \lambda_1, \lambda_2, \ldots, \lambda_K, \sim \text{Cat}(. \mid \lambda_1, \lambda_2, \ldots, \lambda_K)
\]

\[
y_\ell \mid \zeta_\ell = k, m_k, \rho_k \sim f_V(.) \mid m_k, \rho_k), \quad \ell = 1, 2, \ldots, N.
\]

If there are only two von Mises distributions, the general inverse logit function simplifies to the inverse logit and the model only needs one Gaussian process. The inverse logit function, $\psi^{-1}(z)$, is
defined to be $\psi^{-1}(z) = \frac{\exp(z)}{1 + \exp(z)}$. The function is a map from $\mathbb{R}$ to $(0, 1)$. $SvM-p$ reduces down to the following, which we will denote as $SvM-p-2$.

\begin{align*}
m_k &\sim \text{Unif}(\cdot \mid 0, 2\pi), \\
\rho_k &\sim \text{Gamma}(\cdot \mid 1, 1), \\
z &\sim \text{GP}(\cdot \mid 0, \Sigma), \\
\lambda_{1,\ell} &= \psi^{-1}(z_{\ell}), \\
P(\zeta_{\ell} = 1) &= \lambda_{1,\ell}, \\
y_{\ell} \mid \zeta_{\ell} = k; m_k, \rho_k &\sim f_{V}(\cdot \mid m_k, \rho_k),
\end{align*}

(3.17)

We make a few more comments before we discuss the properties of the models. $SvM-c$ and $SvM-p$ can also be thought of in the framework of the Dependent Dirichlet Process or DDP [43]. To understand the DDP, we need to briefly introduce the Dirichlet Process. Given a base distribution, $F_0$, and for $i \in 1, 2, \ldots$, a Dirichlet process assigns weights, $w_i$, to atoms or draws, $\xi_i$, from $F_0$. If $v_i \sim \text{Beta}(1, m)$, then the weight for $\xi_i$ is given by $v_i \prod_{j=1}^{i-1} (1 - v_j)$ in the stick breaking representation [65]. A Dependent Dirichlet Process extends the Dirichlet Process by attaching separate stochastic processes to $v_i$ and $\xi_i$. If a stochastic process is based on location, nearby distributions that use DDP atoms can resemble each other [43]. While we assume a finite number of components, we view $\xi_i$ as the means of a von Mises distributions for our models. Then, in this framework, $SvM-c$ only attaches a stochastic process to $\xi_i$ whereas $SvM-p$ only places a stochastic process on $v_i$. Note that we do not endow stochastic processes to both $\xi_i$ and $v_i$ in this work because such a model might be too rich to depend only on the spatial information.

We also wish to discuss some notation. Because of our convention with $SvM-p-2$, we will denote the number of von Mises distributions after the model if we need to specify $K$. For instance, $SvM-c-3$ indicates the model $SvM-c$ with $K = 3$. The one exception to this guideline is $SvM$, which is $SvM-c$ with $K = 1$ and introduced in the supplementary material (cf, (B.3)).

Finally, despite the focus on the circular case, the framework we introduce can easily be extended into directions of higher dimensions. Based on the discussion in Section 3.2.4, we can plug in the von Mises-Fisher in place of the von Mises distribution. We can also use the higher dimension version of the projected Gaussian process as a prior for the von Mises-Fisher for $SvM-c$. The output from that process is marginally a unit vector, which can serve as the mean parameter for a von Mises-Fisher distribution.
3.3.2 Model properties for spatially dependent von Mises models

To better understand our models, we shall derive the prior circular mean, circular variance, and circular correlation for $SvM$ and $SvM-p-2$. Due to space constraints we do not display the results for $SvM-c$ because one can easily combine the iterated expectation formula with the results for the special case $SvM$ to obtain such results. Due to lack of conjugacy, deriving properties for posterior distributions is difficult. Nonetheless, the following results are useful because practitioners can use them when setting informative priors. While the definition for the circular mean and variance are given in equations (3.7) and (3.8) respectively, we will calculate a notion of correlation proposed by [36]. If $y_\ell$ and $y_{\ell'}$ have means $\alpha_\ell$ and $\alpha_{\ell'}$ respectively, then the correlation is defined to be

$$\text{Corr}(y_\ell, y_\ell') = \frac{\mathbb{E}(\sin(y_\ell - \alpha_\ell)\sin(y_{\ell'} - \alpha_{\ell'}))}{\sqrt{\mathbb{E}(\sin^2(y_\ell - \alpha_\ell))\mathbb{E}(\sin^2(y_{\ell'} - \alpha_{\ell'}))}}. \quad (3.18)$$

These expectations are taken over the interval $[0, 2\pi)$ with respect to some distribution on the circle, $f_c(y_\ell, y_{\ell'})$, or its marginal distributions. As noted by Jammalamadaka and Sarma, this is equivalent to

$$\text{Corr}(y_\ell, y_\ell') = \frac{\mathbb{E}(\cos(y_\ell - \alpha_\ell - (y_{\ell'} - \alpha_{\ell'})) - \cos(y_\ell - \alpha_\ell + y_{\ell'} - \alpha_{\ell'}))}{2\sqrt{\mathbb{E}(\sin^2(y_\ell - \alpha_\ell))\mathbb{E}(\sin^2(y_{\ell'} - \alpha_{\ell'}))}}. \quad (3.19)$$

At a high level, the circular correlation is comparing how tightly $y_\ell - \alpha_\ell$ concentrates around $y_{\ell'} - \alpha_{\ell'}$ versus how tightly $y_\ell - \alpha_\ell$ concentrates around $-(y_{\ell'} - \alpha_{\ell'})$. Then, the correlation defined in this manner has the usual desired properties [36]. For instance, if $y_\ell$ and $y_{\ell'}$ are independent, the correlation is 0.

$SvM$ For $SvM$ specified in (B.3), we will assume that there is a global $\rho$. This is not unreasonable because when we use a hierarchical prior for $\rho_\ell$, we set $\varsigma^2$ to a small value. Then, the circular mean, variance, and correlation are the following.

**Lemma 3.3.1.** If $Y_\ell$ and $Y_{\ell'}$ are generated according to $SvM$ outlined in (B.3) with the random variables associated with them labeled accordingly and $(Z_1, Z_2) \sim \mathcal{N}((\mu_1, \mu_2), \sigma^2\mathbb{I}_2)$ with $\mu_1 = \mu_0 \cos(\alpha_\mu)$ and $\mu_2 = \mu_0 \cos(\alpha_\mu)$, $\mu_0 \in \mathbb{R}^+$, $\alpha_\mu \in [0, 2\pi)$, then

$$\mathbb{E}(Y_\ell) = \alpha_\mu, \quad (3.20)$$

$$\text{Var}(Y_\ell) = 1 - \frac{I_1(\rho)}{I_0(\rho)} \left( \frac{\pi \beta}{2} \right)^{\frac{1}{2}} \exp(-\beta)(I_0(\beta) + I_1(\beta)) \quad \beta = \frac{\mu_0^2}{4\sigma^2}, \quad (3.21)$$

$$\text{Corr}(Y_\ell, Y_{\ell'}) = \frac{\left( \frac{I_1(\rho)}{I_0(\rho)} \right)^2 \left( \mathbb{E}(\cos(m_\ell - m_{\ell'})) - \mathbb{E}(\cos(m_\ell + m_{\ell'} - 2\alpha_\mu)) \right)}{\sqrt{\left( 1 - \frac{I_1(\rho)}{I_0(\rho)} \mathbb{E}(\cos(2m_\ell - \alpha_\mu)) \right)}} \left( 1 - \frac{I_1(\rho)}{I_0(\rho)} \mathbb{E}(\cos(2m_{\ell'} - \alpha_\mu)) \right). \quad (3.22)$$

47
The above lemma reveals the following characteristics of SvM. First, the prior circular mean for $m_\ell$ is $\alpha_\mu$. It is nice that this mean can be set so easily. Based on this, we say that if $\mu_1 = c_1 \mathbb{1}$ and $\mu_2 = c_2 \mathbb{1}$, then SvM is centered at or has mean $\arctan^*(\frac{c_2}{c_1})$. Because this holds for any $\mu_{k,1} = c_{k,1} \mathbb{1}$ and $\mu_{k,2} = c_{k,2} \mathbb{1}$, we might also say that $m_k$ of SvM-c is centered at $\arctan^*(\frac{c_{k,2}}{c_{k,1}})$. Second, the prior circular variance involves multiplying the concentrations of the von Mises’ distribution and mean parameter according to the projected normal distribution. This demonstrates a few ways the Gaussian process can change the prior circular variance without changing the prior circular mean. Recall that $\beta = \frac{\mu_0}{4\sigma^2}$ in (3.21). As $\beta \to 0$, $\text{Var}(Y_\ell) \to 1$ whereas the asymptotics of Bessel functions can be used to show that $\text{Var}(Y_\ell) \to 0$ when $\beta \to \infty$ and $\rho \to \infty$ [72]. The former occurs if $(\mu_1, \mu_2) = (0, 0)$ or $\sigma^2 \to \infty$. As a result, the respective distribution for $m_\ell$ or $Z_1$ and $Z_2$ becomes uniform. On the other hand, $\beta \to \infty$ when $\sigma^2 \to 0$ or $\mu_0 \to \infty$ while $\alpha_\mu$ remains unchanged. In other words, $(Z_1, Z_2)$ becomes more concentrated at $(\mu_1, \mu_2)$ or $P(m_\ell \neq \alpha_\mu) \to 0$. Note however that even if $m_\ell = \alpha_\mu$ with probability 1, but $\rho$ is finite, the circular variance of $Y_\ell$ is that of a von Mises distribution with concentration parameter, $\rho$. If $\rho = 0$, the circular variance is 1 and $Y_\ell$ will be uniformly distributed. This further supports our earlier observation that when SvM is considered in a generative sense, the von Mises distribution adds noise. Finally, it is difficult to compute the exact form of the circular correlation due to the interactions between the terms in the expectation and the projected normal distribution. Fortunately, it is straightforward to compute these values by simulation. Wang and Gelfand briefly explored this in their 2014 paper on projected Gaussian processes.

**SvM-p** We next compute the circular properties of SvM-p. Unlike before, there is no explicit distribution for the inverse logit or general inverse logit transformation of a normally distributed variable. Consequently, we will first derive a formula for any choice of hyperparameters to allow practitioners to understand how their choice will affect these quantities. We do so under the assumption that the von Mises parameters, $m_1, m_2, \ldots, m_K$, and $\rho_1, \rho_2, \ldots, \rho_K$, are fixed. This assumption is a simplification because we can set $m_1, m_2, \rho_1, \rho_2$ to their prior expected values. We then will calculate them for the choices we specified in this chapter for the SvM-p-2 model. Under these assumptions, we have the following lemma.

**Lemma 3.3.2.** If $Y_\ell$ and $Y_\ell'$ are generated according to SvM-p specified in (3.16) with the random variables associated with them labeled accordingly, $\arctan^*$ is defined in (3.1), $\alpha = E(Y_\ell)$, which
is given in (3.23), and
\[ s(\zeta) = 1 - \sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(2(m_k - \alpha)), \]
\[ \mathbb{E}(Y_\ell) = \arctan^* \left( \frac{\sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \sin(m_k)}{\sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(m_k)} \right), \] (3.23)
\[ \text{Var}(Y_\ell) = 1 - \sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(m_k - \alpha), \] (3.24)
\[ \text{Corr}(Y_\ell, Y_{\ell'}) = \frac{1}{s(\zeta_\ell)s(\zeta_{\ell'})} \sum_{k'} \sum_k P(\zeta_\ell = k, \zeta_{\ell'} = k') \frac{I_1(\rho_k)}{I_0(\rho_k)} \frac{I_1(\rho_{k'})}{I_0(\rho_{k'})} \left( \cos(m_k - m_{k'}) - \cos(m_k + m_{k'} - 2\alpha) \right). \] (3.25)

We now discuss the properties in Lemma 3.3.2. In (3.23), \( \text{SVM-p} \)'s mean direction is an average of the components’ mean direction weighted by how concentrated each component is around its mean direction and the probability that an observation belongs to a component. If there are two components that are equally concentrated around their mean direction and equally likely of being selected, we can use trigonometric sum to product formulas to show that \( \alpha \) will be the average of the components’ mean angles. Otherwise, simulation must be used to understand how changing the hyperparameters of the Gaussian process affects \( P(\zeta = k) \) and the mean angle \( \alpha \). The circular variance in (3.24) is similar to the circular variance of \( \text{SVM} \) described in (3.21). From each component, we get the product of how tightly the observation concentrates around that component’s mean direction and how close that component’s mean direction is to the model’s mean direction. The circular variance in (3.24) is also like the circular mean of \( \text{SVM-p} \) in (3.23) because the probability parameters again serves as the weights for a convex combination of each component’s concentrations.

The numerator of the circular correlation in (3.25) is of interest. Each term in the numerator’s sum is comprised of three things. First, it has a comparison between the means concentrating around each other and the difference of one mean from the model mean \( \alpha_2 \) concentrating around its negative counterpart. Next, it holds the product of the observations’ concentration around the mean directions with respect to the observations’ cluster. Finally, it contains the joint probability of the observations’ cluster memberships. It is this probability that will be affected by the spatial information and Gaussian process. At a high level, if observations are "nearby", they are likely to belong to the same cluster. More weight will be placed on the terms coming from the observations belonging to the same cluster. What observations are considered closed is determined by the hyperparameters of the Gaussian process.

\( \text{SVM-p-2} \) We now calculate these properties for the \( \text{SVM-p-2} \) model. Using the notation of the
model (3.17), we will also assume that \( Z, Z_\ell, Z_{\ell'} \sim N(0, 1) \) and \( s = \text{Cov}(Z_\ell, Z_{\ell'}) \). One technical issue arising in computing these quantities is that \( P(\zeta = k) = \mathbb{E}(\psi^{-1}(Z_\ell)) \), which is intractable. Instead, we will bound any expectation involving the logistic function with the functions, \( f(z_\ell) \) and \( g(z_\ell) \). We define \( f(z_\ell) = g(z_\ell) + \frac{1}{2} \), where \( g(z_\ell) \) is defined as following:

\[
g(z_\ell) = \begin{cases} 
-\frac{1}{2} & z_\ell < -z_e \\
\frac{1}{2} & z_\ell \in [-z_e, z_e] \\
\frac{1}{2} & z_\ell > z_e, 
\end{cases}
\] (3.26)

We will additionally assume that \( z_e \in (0, 2] \) to ensure that \( f(z_\ell) < \psi^{-1}(z_\ell) \) for \( z < 0 \) and \( f(z_\ell) > \psi^{-1}(z_\ell) \) for \( z > 0 \). Under such assumptions, \( \mathbb{E}(f(Z)) = \mathbb{E}(\psi^{-1}(Z)) \). As a result, these functions, \( f(z_\ell) \) and \( g(z_\ell) \), are not a bad choice.

We can also use these functions to prove that \( \mathbb{E} (\psi^{-1}(Z)) = \frac{1}{2} \) under our assumptions. This fact and Lemma 3.3.2 then gives us the following lemma.

**Lemma 3.3.3.** If \( Y_\ell \) and \( Y_{\ell'} \) are generated according to SvM outlined in (3.17) with the random variables associated with them labeled accordingly, \( Z \sim N(0, 1) \), \( (Z_\ell, Z_{\ell'}) \sim N(0, \Sigma) \), \( z_e \in (0, 2] \), and \( \alpha = \mathbb{E}(Y_\ell) \), which is given in (3.27), then

\[
\mathbb{E}(Y_\ell) = \arctan^* \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \sin(m_1) + \frac{I_1(\rho_2)}{I_0(\rho_2)} \sin(m_2) \right) 
\] (3.27)

\[
\text{Var}(Y_\ell) = 1 - \frac{1}{2} \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \cos(m_1 - m_2) + \frac{I_1(\rho_2)}{I_0(\rho_2)} \cos(m_2 - m_1) \right) 
\] (3.28)

\[
\text{Corr}(Y_\ell, Y_{\ell'}) = \frac{2\mathbb{E}(\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha))}{1 - \frac{1}{2} \left( \frac{I_2(\rho_1)}{I_0(\rho_1)} \cos(2m_1 - m_2) + \frac{I_2(\rho_2)}{I_0(\rho_2)} \cos(2m_2 - m_1) \right)}. 
\] (3.29)

We make the following comments based on the lemma above. First, because both component’s are equally favored, \( m_1, m_2, \rho_1, \) and \( \rho_2 \) for the prior mean and variance become more important. The \( \text{SvM-p-2} \) model’s prior mean direction is completely determined by these parameters. For the model’s prior circular variance, the concentration term is the average of the two components’ concentration terms, which underscores the importance of \( \rho_1 \) and \( \rho_2 \). If we want the Gaussian process to have more of an effect on the prior mean or variance, it is necessary to change the mean for the Gaussian process. It can be shown that \( \mathbb{E}(f(Z)) = \mathbb{E}(\psi^{-1}(Z)) \) for \( Z \) distributed according to any normal distribution with mean 0. Next, \( \mathbb{E}(\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha)) \) in (3.29) can be expanded and written as a function of \( \mathbb{E}(\psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'})) \). Doing so shows that the numerator again represents the difference between the concentration of two observations belonging to the
same component and the concentration of two observations belonging to different components. It is then possible to bound $E \left( \psi^{-1} (z_{\ell}) \psi^{-1} (z'_{\ell}) \right)$ with $E \left( f(z_{\ell}) f(z'_{\ell}) \right)$. Due to the length and technical aspects, we leave the details of this derivation and the expansion to the supplementary material.

3.4 Posterior inference

We now describe how to fit our models. Fitting $SvM-c$ can be difficult because of the underlying non-Euclidean geometry of its parameter space. We will first describe the basic sampling schemes based on the Hamiltonian MCMC approach, and then discuss techniques for exploiting the geometric structures via a suitable application of elliptical slice sampling.

**Sampling algorithms** We use the following approaches to sample from our models. For $SvM-p$, we use Hamiltonian Monte Carlo (HMC) to sample from the following posterior:

$$\prod_{\ell=1}^{N} \left( \sum_{k=1}^{K} \lambda_k f_V (y_{\ell} \mid m_k, \rho_k) \right) \prod_{k=1}^{K} \Gamma (\rho_k \mid 1, 1) \prod_{k=1}^{K} N \left( \tilde{z}_k \mid 0, \mathbb{I}_{N \times N} \right).$$

Here, $\tilde{z}_k$ is the vector such that if $\Sigma = LL^T$, then $z_k = L \tilde{z}_k$. Further, note that the probability for a single observation $y_{\ell}$ is the following:

$$p(y_{\ell} \mid \tilde{z}_{., \ell}, m, \rho) = \sum_k \lambda_{k, \ell} f_V (y_{\ell} \mid m_k, \rho_k).$$

In other words, we marginalize out the label. As a result, we are using HMC to sample for $m_1, m_2, \ldots, m_K; \rho_1, \rho_2, \ldots, \rho_K$; and $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{K-1}$ according to the non-centered parametrization.

Then, the gradients for the log likelihood and thus the update for the momentum vector are given below.

- For $k = 1, 2, \ldots, K - 1$ and $\ell = 1, 2, \ldots, N$, the gradient for $\tilde{z}_{k, \ell}$ is the following:

$$- \frac{\partial U}{\partial q_{\ell}} (q(t)) = \sum_{\ell'} L_{\ell', \ell} \lambda_{k, \ell'} \frac{f_V (y_{\ell} \mid m_k, \rho_k) - \sum_{k'=1}^{K} \lambda_{k', \ell'} f_V (y_{\ell} \mid m_{k'}, \rho_{k'})}{p(y'_{\ell} \mid \tilde{z}_{., \ell'}, m, \rho)} - \tilde{z}_{\ell}. \quad (3.30)$$

In the two parameter case, this reduces to the following:

$$- \frac{\partial U}{\partial q_{\ell}} (q(t)) = \sum_{\ell'} \lambda_{1, \ell'} (1 - \lambda_{1, \ell'}) \frac{f_V (y_{\ell} \mid m_1, \rho_1) - f_V (y_{\ell} \mid m_2, \rho_2)}{p(y_{\ell} \mid z_{\ell}, m_1, m_2, \rho_1, \rho_2)} L_{\ell', \ell} - \tilde{z}_{\ell}. \quad (3.31)$$

51
• For \( k = 1, 2, \ldots, K \), the gradient for \( m_k \) is the following:

\[
- \frac{\partial U}{\partial q_k}(q(t)) = \sum_v f_v(y_v \mid m_k, \rho_k) \sin(y_v - m_k).
\] (3.32)

• For \( k = 1, 2, \ldots, K \), the gradient for \( \rho_k \) is the following:

\[
- \frac{\partial U}{\partial q_k}(q(t)) = \sum_v \frac{f_v(y_v \mid m_k, \rho_k)}{p(y_v \mid \tilde{z}, e, m, \rho)} \left( \cos(y_v - m_k) - \frac{I_{1-1}(\rho_k)}{I_0(\rho_k)} \right) - 1.
\] (3.33)

On the other hand, for \( \text{SVM-c} \), we use the following blocked Gibbs approach:

• For \( \ell = 1, 2, \ldots, N \), sample \( \zeta_\ell \) given \( y_\ell; z_{1,1,\ell}, z_{1,2,\ell}, z_{2,1,\ell}, z_{2,2,\ell}, \ldots, z_{K,1,\ell}, z_{K,2,\ell}; \varphi_{1,\ell}, \varphi_{2,\ell}, \ldots, \varphi_{K,\ell}; \) and \( \nu_1, \nu_2, \ldots, \nu_K \) according to the following probability:

\[
p(\zeta_\ell = k \mid y_\ell; z_{1,1,\ell}, z_{1,2,\ell}; \varphi_\ell; \nu_\ell) \propto \lambda_k f_V(y_n \mid m_k, \rho_k, \ell).
\]

• For \( k = 1, 2, \ldots, K \), sample \( (z_{k,1} - \mu_{k,1}, z_{k,2} - \mu_{k,2}) \) given \( y; \zeta_1, \zeta_2, \ldots, \zeta_N; \varphi_1, \varphi_2, \ldots, \varphi_K; \) and \( \nu_1, \nu_2, \ldots, \nu_K \) using the elliptical slice sampling outlined in Algorithm 2 [50]. If we use the terminology of the paper, the normal prior and the likelihood that we are sampling are

\[
\mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{k,1} & 0 \\ 0 & \Sigma_{k,2} \end{pmatrix} \right) \) and \( \prod_\ell f_V(y_n \mid m_{k,\ell}, \rho_{k,\ell})^{I(\zeta_\ell = k)} \) respectively.

• Sample \( \varphi_1, \varphi_2, \ldots, \varphi_K \) and \( \nu_1, \nu_2, \ldots, \nu_K \) given \( y; \zeta_1, \zeta_2, \ldots, \zeta_N; \) and \( z_{1,1}, z_{1,2}, \ldots, z_{K,1}, z_{K,2} \) using HMC and the following probability:

\[
p(\varphi_1, \varphi_2, \ldots, \varphi_K, \nu_1, \nu_2, \ldots, \nu_K \mid y; \zeta_1, \zeta_2, \ldots, \zeta_N; z_{1,1}, z_{1,2}, \ldots, z_{K,1}, z_{K,2}) \propto \prod_\ell \left( f_V(y_\ell \mid m_{k,\ell}, \rho_{k,\ell})^{I(\zeta_\ell = k)} \right) \prod_k \left( \prod_\ell \mathcal{N}(\varphi_{k,\ell} \mid \nu_k, \tau^2) \right) \mathcal{N}(\nu_k \mid 0, \tau^2).
\]

Unlike \( \text{SVM-p} \), we sample for the labels, \( \zeta_1, \zeta_2, \ldots, \zeta_N \), because the mixing appeared to be better according to the trace plots.

To help the samplers, we used initial values obtained via a regularized version of Expectation Maximization algorithm derived from \( \text{SVM-c} \) and \( \text{SVM-p} \). We leave the details for these algorithms to the supplementary material.

**Exploiting geometric structure** To sample \( \text{SVM-c} \) in an efficient manner, we need to sample for the mean angles while respecting the spatial information, which is embedded in the Gaussian process’ covariance matrix. HMC has trouble doing both even if we change the model parametrization. In one parametrization, the mean angles can be sampled using HMC, but at the cost of having to invert
Figure 3.2: Plots showing the elliptical slice sampler’s proposal ellipse for different locations (shown as circles) and means for the Gaussian prior (shown as the squares), but the same elliptical proposal point (shown as the triangle). These ellipses represent the set of angles that can be proposed by the elliptical slice sampler. The arrow demonstrates that the locations have the same angle in polar coordinates, but different radii.

the covariance matrix. An alternative parametrization allows HMC to propose moves according to the covariance matrix without inverting it. However, these moves marginally lie in $\mathbb{R}^2$ and may not result in new angles being sampled even if the proposal is accepted. For example, a marginal move along a ray from the origin might be accepted even though $m_\ell$ remains the same. Note that this is not the case for $SvM-p$ because the Gaussian processes are mapped to probabilities via the general inverse logit function. Hence, each marginal move represents a different mixing probability for each observation. More discussion of why HMC struggles for $SvM-c$ and not $SvM-p$ and the exact parametrization discussed are given in the supplementary material.

We turn to elliptical slice sampling instead [50]. This technique allows us to draw samples for a random variable distributed according to a Gaussian prior with zero mean and any likelihood. For $SvM-c$, we use the elliptical slice sampler to sample for $(Z_{k,1} - \mu_{k,1}, Z_{k,2} - \mu_{k,2}) \mid \varphi_k, \nu_k, \zeta_k$ for $k \in 1, 2, \ldots K$. In this case, the Gaussian prior has covariance $I_2 \otimes \Sigma$. Following the notation of the paper, we let $f$ represent the current location of $(Z_{k,1} - \mu_{k,1}, Z_{k,2} - \mu_{k,2})$. If we use the transformation specified in (3.1) on $f + (\mu_1, \mu_2)$ to get $m$, then the likelihood is defined as follows:

$$
L(y \mid m_k, \varphi_k, \nu_k, \zeta_k) = \prod_{\ell} (f_{\psi} (y_\ell \mid m_\ell, \rho_\ell))^{\Pi (\zeta_\ell = k)}.
$$

Note that even if $y_\ell$ is not included in the likelihood because $\zeta_\ell \neq k$, we still propose $Z_{k,1,\ell}$ and
$Z_{k,2,t}$. This will prove useful for predictive purposes.

As seen in Figure 3.2, the elliptical slice sampler is marginally a sensible way to explore the space if $\mu_1 = 0$ and $\mu_2 = 0$. We can then explore the entire space of $m_\ell$ with each run of the slice sampler. Unfortunately, if we set $\mu_1$ and $\mu_2$ to be 0, $m_\ell$ will be marginally distributed according to a uniform distribution. By not setting $\mu_1$ and $\mu_2$ to be 0, Figure 3.2 shows us that not only can the sampler potentially explore a subset of random angles, but also the subset might depend on its current location in $\mathbb{R}^2$ and the point used to determine the ellipse. This might affect the next draw because for the same value of $m_\ell$, the range of the next $m_\ell$ may be different. Still, even with this limitation, the elliptical slice sampler allows us to proposes a valid move in the space of $m$ that respects the covariance information from the Gaussian process. Due to it being hard to characterize the dependence of the next draw on the location, we thin our MCMC chain to mitigate this issue. Luckily, the elliptical slice sampler runs reasonably quickly so this does not impose too much of an additional computational burden.

3.5 Simulations and results

In this section we shall present a simulation study for the introduced models and then an application to the analysis of the income proportion data set.

3.5.1 Simulation study

Simulation overview The simulation study is conducted with following goals in mind. First, we wanted evidence that if we used our models for inference, our models could correctly recover the model parameters for data generated according to their respective models. Next, with this menagerie of simulated data, we could examine how these models behave in mis-specified settings and how to address model selection. Then, while fitting these models, we might notice inefficiencies in our sampling scheme and develop better approaches. Finally, inspecting the results might yield additional insights about the behaviors of these models.

To that end, we used a uniform distribution to generate 500 "locations" in $\Delta^2$. We created observations in six different ways. First, we generated observations using von Mises($\pi$, 5). Next, we created them using a mixture of von Mises($\frac{\pi}{2}$, 5) and von Mises($\frac{3\pi}{2}$, 10) with mixing probability of 0.3 and 0.7 respectively. As discussed in the supplementary material, these models are the $iV$ and $iVM$ models respectively. While these two methods did not use the spatial information, the remaining methods do. Third, we simulated observations according to $S\nu M$ specified in Model (B.3) with a prior mean of $\pi$ for all locations. Fourth, we created them according to $S\nu M-c$ specified in Model (3.15). One component’s mean was centered at $\frac{\pi}{2}$ and the other component’s mean at $\frac{3\pi}{2}$. 

54
There was equal probability of using either component. Next, we used the \textit{SvM-p} model, \eqref{eq:3.16}, to simulate observations. We again used means of $\frac{\pi}{2}$ and $\frac{3\pi}{2}$. Finally, we simulated according to \textit{SvM} specified in Model \eqref{eq:B.3} with a prior mean of 0 for all locations. This will test whether our model can handle data that appears to be separated due to where we set zero, but is actually connected.

**Simulation model fitting** We fit the \textit{SvM}, \textit{SvM-c}, and \textit{SvM-p} models to these simulated observations using the approaches outlined in the previous section. Because there was at most two components, we only fit the two component versions of \textit{SvM-c} and \textit{SvM-p} to our simulated observations. Then, for the Gaussian processes, we used the squared exponential kernel with $\omega = 0.1$. For \textit{SvM-p}, we set $\sigma = 1$ and $\mu = 0$. Meanwhile, for \textit{SvM} and \textit{SvM-c} model, we set $\sigma = 0.5$. We set $\mu = (-1, 0)$ for \textit{SvM} and $\mu_1 = (0, 1)$ and $\mu_2 = (0, -1)$ for \textit{SvM-c}. These values correspond to $\pi$, $\frac{\pi}{2}$, and $\frac{3\pi}{2}$ respectively. We set $\varsigma = 0.05$ and $\tau = 5$ for \textit{SvM} and \textit{SvM-c} concentration parameter’s hierarchical prior. Our goal for these choices of parameters was to capture a larger range of probabilities for \textit{SvM-p} whereas we wanted components to be more distinct for \textit{SvM} and \textit{SvM-c} models.

For posterior inference, any step involving HMC was implemented in Stan \cite{c1}. Otherwise, we implemented our approach in R. Using this code\footnote{Our code can be found at \url{https://github.com/rayleigh/housing_price}.}, we ran four chains and 2000 iterations of the HMC sampler for \textit{SvM-p} with 1000 iterations as burn-in. Meanwhile, for \textit{SvM} and \textit{SvM-c} sampler, we ran four chains and 10000 iterations, using every fifth iteration after the first 5000 iterations. On a cluster using 2x 3.0 GHz Intel Xeon Gold 6154 as its processor and with 192 GB as its RAM, \textit{SvM-c} took between four and a half to eight and a half hours whereas \textit{SvM-p} took between 8 and 20 minutes. We checked for convergence by primarily looking at trace plots of parameters and examining Rhat values for non-circular values as computed by Stan \cite{c2}.

**Simulation model findings** These are a number of interesting results to report from these experiments. First, as exemplified by Figure 3.3 and other such figures in the appendix, the models appeared to correctly capture the mean surfaces for data simulated from their respective models. For example, the fitted mean surface for \textit{SvM} matches the simulated mean surface, even when the simulated surface is centered at 0. While the numerical summaries of the random variables were slightly off, this could be due to the posterior credible intervals being averaged across all locations. Next, homogeneous models may have trouble capturing heterogeneous patterns. For instance, the fitted mean surface from \textit{SvM} tries to match the observations as much as possible in the case of \textit{SvM-p} or essentially gives up and becomes a uniform distribution in the case of \textit{SvM-c}. It does not recognize the simulated mean pattern of either model. Finally, the results demonstrate that \textit{SvM} and \textit{SvM-c} are in a different class of models compared to the class of models \textit{SvM-p} belongs to. As seen in Figure 3.3, \textit{SvM-p} tries to put the von Mises distributions at the two components’ overall mean with small concentration parameters in the case of \textit{SvM-c}. The fitted model is trying to make up for the variation in the observations by returning two diffuse von Mises distribution. A similar
Figure 3.3: The right three plots show the mean component fitted by different models for random directions simulated according to $SVM-c$ in (3.15). The fitted mean components for the different models are shown in colored asterisks with the simulated random directions as circles and the simulated means as grey rhombuses. The leftmost plot is just the simulated data and the mean surface used to generate that data. The x-y axis show the locations in the first two dimensions of a three dimensional simplex; the z-axis represent the direction.

Table 3.2: Circular mean posterior values and 95% credible intervals for the parameters in von Mises distribution and mixing probability are shown for the simulation data in Figure 3.3. Parameters with bars over them were averaged across all locations.

<table>
<thead>
<tr>
<th>Model</th>
<th>$m_1$</th>
<th>$m_2$</th>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>$m_1 = 1.36 (0.41, 2.16)$</td>
<td>$m_2 = 4.85 (3.53, 5.64)$</td>
<td>$\rho_1 = 3.01 (2.56, 3.50)$</td>
<td>$\rho_2 = 8.00 (6.98, 9.00)$</td>
<td>$\lambda_1 = 0.50$</td>
<td>$\lambda_2 = 0.50$</td>
</tr>
<tr>
<td>SVM cf. (B.3)</td>
<td>$m = 3.15 (1.93, 4.35)$</td>
<td>$\rho = 0.01 (0.00, 0.04)$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SVM-c cf. (3.15)</td>
<td>$m_1 = 1.42 (0.90, 1.97)$</td>
<td>$m_2 = 4.79 (4.38, 5.20)$</td>
<td>$\rho_1 = 2.79 (2.17, 3.52)$</td>
<td>$\rho_2 = 8.01 (6.11, 10.52)$</td>
<td>$\lambda_1 = 0.54 (0.49, 0.58)$</td>
<td>$\lambda_2 = 0.46 (0.42, 0.51)$</td>
</tr>
<tr>
<td>SVM-p cf. (3.16)</td>
<td>$m_1 = 1.38 (1.24, 1.51)$</td>
<td>$m_2 = 4.77 (4.68, 4.86)$</td>
<td>$\rho_1 = 1.44 (1.07, 1.91)$</td>
<td>$\rho_2 = 4.08 (2.75, 5.55)$</td>
<td>$\lambda_1 = 0.58 (0.32, 0.82)$</td>
<td>$\lambda_2 = 0.42 (0.18, 0.68)$</td>
</tr>
</tbody>
</table>

result happens in the $SVM$ case, except that the two von Mises are placed at an "lower" and "upper" mean. In contrast, $SVM-c$ correctly recovers the simulated mean surfaces in the $SVM-p$ case, but it can only roughly capture the average probability of each surface. Interestingly enough, in the $SVM$ case, $SVM-c$ matches one fitted mean surface to the simulated mean surface and essentially "zeroes" out the other surface.

Simulation model selection We decided to compute the posterior predictive probability for another
Table 3.3: The log of the posterior predictive probability given in Equation (3.34) for 50 draws averaged across posterior draws and then posterior predictive draws.

<table>
<thead>
<tr>
<th></th>
<th>iV</th>
<th>iVM</th>
<th>SvM Mean</th>
<th>π</th>
<th>SvM-c</th>
<th>SvM-p</th>
<th>SvM Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>iV</td>
<td>-39.85</td>
<td>-78.20</td>
<td>-68.08</td>
<td>-91.99</td>
<td>-92.05</td>
<td>-89.20*</td>
<td></td>
</tr>
<tr>
<td>iVM</td>
<td>-39.92</td>
<td>-47.60</td>
<td>-68.24</td>
<td>-86.47</td>
<td>-61.80</td>
<td>-85.69</td>
<td></td>
</tr>
<tr>
<td>SvM</td>
<td>-42.42</td>
<td>-84.28</td>
<td>-59.34</td>
<td>-92.43</td>
<td>-97.86</td>
<td>-63.49</td>
<td></td>
</tr>
<tr>
<td>SvM-c</td>
<td>-45.78</td>
<td>-56.79</td>
<td>-62.19</td>
<td>-73.40</td>
<td>-70.90</td>
<td>-60.67</td>
<td></td>
</tr>
<tr>
<td>SvM-p</td>
<td>-39.98</td>
<td>-49.60</td>
<td>-66.61</td>
<td>-90.03</td>
<td>-62.78</td>
<td>-73.54</td>
<td></td>
</tr>
</tbody>
</table>

50 random locations and their observations simulated according to the different scenarios in order to model select. Let $x^*$ represent the withheld locations, $y^*$ the withheld data, $\theta$ a posterior draw for the parameters based on $x$ and $y$, and $\theta^*$ a draw for the parameters for $x^*$ and $y^*$. The posterior predictive probability is given as follows.

$$p(y^* \mid x, x^*, y) := \int \int p(y^* \mid \theta^*) p(\theta^* \mid \theta, x, x^*) p(\theta \mid x, y) d\theta^* d\theta.$$  (3.34)

We discuss how to compute this probability in the supplementary material because it is straightforward to compute $p(y^* \mid \theta^*)$ according to our models defined in (3.15) and (3.16). It is also simple to sample from $p(\theta^* \mid \theta, x, x^*)$ due to the conditional formula for the Multivariate normal distribution to generate these draws from the Gaussian processes. We take advantage of this fact by sampling 100 times from $p(\theta^* \mid \theta, x, x^*)$ to create our posterior predictive draws. To ensure that both the posterior draws and the posterior predictive draws speak equally, we first averaged across posterior draws. We then averaged these means across the posterior predictive draws. This gives us a sense how likely our posterior is on withheld data.

Table 3.3 shows the results from doing so. There are a few interesting things to note. First, the posterior predictive probability confirms that the homogeneous models perform poorly in any heterogeneous scenarios. Next, the posterior predictive probability calculated after fitting $SvM-p$ is higher when the means of the von Mises used to generate the data are constant. On the other hand, the posterior predictive probability calculated after fitting the $SvM$ and $SvM-c$ models is higher when the von Mises components’ means are spatially correlated. We might expect the former because spatially correlated observations might be overfitted to the data. For the latter, $SvM-p$ might have difficulty in capturing the variability and thus cannot accurately predict the
Figure 3.4: Plots showing the change in income proportion for census tract 6037920336 and its neighboring tract on the left and the random direction of the movements for 2003-2004 by census tract on the right. The center of the census tract in longitude and latitude is used to represent the census tract.

next value. Finally, while we do not report these cases, there are scenarios in which \( SVM-c \) can perform poorly according to the predictive posterior probability. When we plot them, we see that the worst performing observations lie between the fitted mean surface. Even though the fitted mean surface and concentration parameter may be mostly correct for the observed data, the model is being penalized for being "over-certain" about the withheld data. This means that if we desire well-separated components, the observations must also be well-separated. Even with this caveat, we use the posterior predictive probability to model select because it allows us to compare models that are different in their approaches and generally picks the model used to generate the data in simulation.

3.5.2 Data analysis

Data overview We now fit the proposed models to the random directions observed in the year to year income proportion changes in the Home Mortgage Disclosure Act (HMDA) data for each census tract of Los Angeles County. While HMDA data is publicly available, the dataset we worked with is not because it is fused with data purchased from a private company. For this chapter, there are three income categories: $0 to $100,000, $100,000 to $200,000, and greater than $200,000. We choose to examine these proportions because the number of mortgages recorded in a year differ per census tracts. Analyzing proportions allows us to potentially ignore the biases that might arise from these differences. We also assume that these proportions observed are the true income proportions for a census tract. This assumption may not be too unreasonable because people are likely to move into neighborhoods with demographic characteristics similar to their own.
To model how these income proportions change, we also decided to use the proportions as locations because as seen in Figure 3.4, it can be challenging to do so otherwise. If we examine how the income proportions evolve for census tract 6037920336 and its adjacent tracts, we see that the income proportion for the top category change in a similar manner. However, the proportions for the lower two income categories only mostly shift in a similar manner after 2005. Further, there is one adjacent census tract that diverges from the other tracts for the second income category during this time period. As a result, it is hard to imagine a model that only uses census tract information because it might need to borrow strength for one income category, but potentially ignore it for other categories. If we then extract the random direction using the procedure discussed in the next paragraph, Figure 3.4 also shows that it might not be helpful to model these directions using census tracts. Even though the random direction is not uniformly distributed, the pattern appears to be the same across the different census tracts.

**Random direction** Because we have the income proportion for a census tract for a given year and the next year, we can extract the random direction using the following procedure. Let \(x_1\) be the income proportion for one year and \(x_2\) be the proportion for the next. Set \(\theta_1\) and \(\phi_1\) to be the spherical coordinates for \(\sqrt{x_1}\). Then, define \(O_p\) to be the following matrix:

\[
\begin{pmatrix}
\cos \theta_1 \cos \phi_1 & -\sin \phi_1 & \sin \theta_1 \cos \phi_1 \\
\cos \theta_1 \sin \phi_1 & \cos \phi_1 & \sin \theta_1 \sin \phi_1 \\
-\sin \theta_1 & 0 & \cos \theta_1
\end{pmatrix}
\]

This matrix transforms \((0, 0, 1)\) to \(x_1\). It is orthogonal because the first column is the spherical coordinates for \(\theta_1 + \frac{\pi}{2}\) and the second column is the spherical coordinates for \(\theta_1 = \frac{\pi}{2}\) and \(\phi_1 + \frac{\pi}{2}\). Its inverse allows us to define a spherical coordinate system based on \(x_1\). If \(O_p x_2' = x_2\) and \(\theta_2'\) and \(\phi_2'\) represent the spherical coordinates of \(x_2'\), then \(\phi_2'\) is the random direction and \(\theta_2'\) represents how "far" \(x_1\) goes in that random direction.

This random direction has some nice properties. Even if \(x_1\) is zero for any proportions, the random direction can still be extracted and examined. It is also interpretable. In spherical coordinates, changing \(\phi_2'\) for a fixed \(\theta_2'\) only affects the first two coordinates. Due to the transformation defined in the previous paragraph, a change in the first two coordinates changes how the first two columns of the matrix that transforms \((0, 0, 1)\) to \(x_1\) are weighted. To understand this random direction, we need to comprehend what the first two columns represent and how the random direction interacts with these columns. Because the first column represents the spherical coordinates for \((\theta_1 + \frac{\pi}{2}, \phi)\), the simplest way to understand this column is to compare \(\theta_1 = 0\) and \(\theta_1 = \frac{\pi}{2}\). As \(\theta_1 = 0\) corresponds to \((0, 0, 1)\) and \(\theta_1 = (\cos(\phi), \sin(\phi), 0)\) for some \(\phi \in [0, 2\pi]\), this first column represents a push away from the third income category. On the other hand, if we use a similar logic to compare \(\theta = \frac{\pi}{2}\) and \(\phi_1 = 0\) against \(\theta = \frac{\pi}{2}\) and \(\phi_1 = \frac{\pi}{2}\), then we see that the second column represents a pull toward
the second income category. Because the first and second coordinate include \( \cos(\phi'_2) \) and \( \sin(\phi'_2) \) respectively, we will examine 0, \( \frac{\pi}{2} \), \( \pi \), and \( \frac{3\pi}{2} \). Note that at 0 and \( \pi \), the first coordinate will be 1 and \(-1\) and the second will be zero by definition. A random direction of 0 represent a push away from the third income category whereas \( \pi \) represents a pull toward. The reverse is true for \( \frac{\pi}{2} \) and \( \frac{3\pi}{2} \). We can say that a random direction of \( \frac{\pi}{2} \) and \( \frac{3\pi}{2} \) represent a pull toward and push away from the second income category. It is the opposite because the first column is a push away from a category whereas the second column is a pull toward a category.

We fitted our models to these random directions with one additional pre-processing step. We removed duplicated directions so that each location has at most one observation. This was done because duplicated directions at the same location should happen with probability zero according to our model and they can be easily identified. Indeed, their properties are further discussed in the supplementary material. As a result of this step, the number of observations are reduced from between 2302 to 2356 per year to between 1704 to 2164 per year. However, while we used the same priors that we utilized for simulation, it was still unclear what to set the hyperparameters for the squared exponential kernels. To pick, we used the same hyperparameters as simulation as a baseline. We then compared it against \( \sigma = 0.25, 1 \), and 2 and \( \omega = 0.05, 0.15, 0.2 \), and 0.25 for \( \text{SVM-c} \) and \( \sigma = 0.25, 0.5 \), and 2 and \( \omega = 0.05, 0.15, 0.2 \), and 0.25 for \( \text{SVM-p} \).

**Hyperparameter and model selection** Some of our findings from these experiments can be seen in Figure 3.5. If we examine \( \text{SVM-p} \) first, we see that while the overall mixing probability averaged across all locations remain mostly the same, the credible intervals averaged across all locations expands as \( \sigma \) increases or \( \omega \) decreases. This is to be expected because increasing \( \sigma \) allows for a wider range of values for the Gaussian process whereas decreasing \( \omega \) means the mixing probability is more sensitive to the observed direction at each location. Next, if we examine \( \text{SVM-c} \), changing the hyperparameters does not appear to affect the first mean averaged by location between 1995 and 2005. However, there is greater stability in the second mean averaged by location when \( \omega \) is

Figure 3.5: Plots showing the mean direction and credible intervals averaged over all locations for various hyperparameters for \( \text{SVM-c} \) compared to \( \omega = 0.1 \) and \( \sigma = 0.5 \) on the left and showing the probability and credible intervals averaged over all locations for various hyperparameters for \( \text{SVM-p} \) compared to \( \omega = 0.1 \) and \( \sigma = 1 \) on the right. The value for \( \omega \) is given first.
changed, but not when $\sigma$ is altered. Indeed, the latter only appears to have stability before 1998. Still, there are differences in the mean surfaces hidden by this statistic. If we focus on the results from 1996 to 1997, we see that while the second mean surface is mostly similar when we change $\sigma$, the first mean surface varies more when $\sigma$ increases. For instance, the surface is mostly flat when $\sigma = 0.25$, but the surface appears to be a series of thinly connected pillars that span zero to $2\pi$ when $\sigma = 2$. This makes sense that neighboring values on the mean surface can vary drastically when $\sigma$ grows. On the other hand, the overall shape of the mean surface in 1996 mostly remains the same as we increase $\omega$ from 0.05 to 0.25. This is again particularly true for the second mean surface. However, the first mean surface becomes smoother as $\omega$ rises. This suggests that for this year, neighboring points are similar in value and the mean surface must reflect this fact.

After running these experiments, we selected the hyperparameters that had the average highest log posterior predictive probability across all years. Here, we take the average with respect to the number of observations. For $SvM-c$, this was $\sigma = 0.5$ and $\omega = 0.2$. The hyperparameters selected for $SvM-p$ was $\sigma = 2$ and $\omega = 0.1$. We then used these results to explore the choice of kernel. The other kernels were the Matern with three halves and five halves degrees of freedom. The Matern with one half degree of freedom was not compared because previous experiments demonstrated that kernels leading to results that are less smooth did not perform as well and the Matern kernel with one half is such a kernel [59]. For $SvM-p$, the results remained similar compared to the results from the optimal squared exponential kernel model. Indeed, the posterior predictive probability is similar between these models. Meanwhile, choosing the Matern kernels or the optimal squared exponential kernel for $SvM-c$ result in similar mean directions averaged by location. However, if we again look at the mean surface from 1996 to 1997, we notice differences. While the lower mean surface is similar from the Matern kernel with five halves degrees of freedom and squared exponential kernel, the surface is more dispersed for the Matern kernel with three half degrees of freedom. However, for non-lower income neighborhoods. the top mean surface is more scattered for the Matern kernel than for the squared exponential kernel. Interestingly enough, the surface from the Matern kernel with five halves degrees of freedom is more diffuse even though it is smoother than the surface from the Matern kernel with three halves degrees of freedom. One other key difference is that the $SvM-c$ models with the Matern kernels performed worse than the same model with the squared exponential kernel. This suggests that while the mean surface for the random direction might be sensitive to the choice of kernel, regions that are likely to move in the same direction are not sensitive.

Finally, we also wanted to compare the results from the square exponential hyperparameter experiments against $SvM$ and $SvM-c$ using the hyperparameters selected for $SvM-c$ and $SvM-p$ with the hyperparameters selected according to $SvM-p$. Note that for years 1991-1993, 1996-1997, 1998-2000, 2005-2010, we used a strong von Mises prior on the mean component to force identifiability for $SvM-p$. Otherwise, we used an uniform prior on the mean component. Further,
Figure 3.6: Plots showing the observed random direction not withheld and the fitted mean surface of the model selected by the posterior predictive log probability in (3.34). The front axis represents the proportion in the first income category and the side axis represents the proportion in the second income category. This is reversed in the alternative view. Regardless, the up-down axis represents the direction. The start of the arrow indicates a value of zero whereas the end indicates a value of $2\pi$ for angles and 1 for proportions.

because of extreme estimates in the concentration parameters, we also ran a version of $SvM\text{-}c\text{-}3$ with the concentration parameter constrained to be between 1e-4 and 20. According to the log posterior predictive probability given in (3.34), $SvM\text{-}c\text{-}2$ perform the best for all years except 1991-1994, 2003-2005, 2006-2007, 2008-2010, and potentially 2008-2009. For those years, $SvM\text{-}c\text{-}3$ or the constrained version performs the best. However, the posterior predictive probability for $SvM\text{-}c\text{-}3$ is close to that of $SvM\text{-}p\text{-}3$ for 2006-2007 and of $SvM\text{-}c\text{-}2$ for 2008-2009. This suggests that the average observed change in income proportion is better explained by one of two random direction surfaces that varies depending on the income proportion proportions. Further, it tells us that having two components is sufficient in most cases.

Findings and interpretation Based on the Table B.1 and Figure B.3 in the supplementary material, the fitted mean surface can be grouped into four phases: 1990-1991, 1991-1993, 1993-2003, 2003-2010. These phases correspond to the early 1990 recession, the transition from President George
HW Bush to President Bill Clinton, the economic boom in the 1990s, and the recovery from the dot com bubble with the subsequent housing market crash respectively. Representative examples of each phase can be found in Figure 3.6. As an example of the stories these surfaces can tell us, consider 2005 to 2006 and 2006 to 2007. Like the two years before 2005, there is one surface that is an curved, half spiral increasing from $\frac{\pi}{2}$ to around $\pi$ if we follow it from neighborhoods with income proportions largely below $100,000 to neighborhoods with income proportions largely between $100,000 and $200,000 and then to neighborhoods with income proportions largely greater than $200,000. This suggests that for these years, the income distributions for all neighborhoods are being pulled up a category. However, there is another surface that while similar for neighborhoods of lower income, is centered around zero for neighborhoods with income proportions in the second and third categories. This suggests that there already is a push away from the third income category for these neighborhoods, two years before the housing market crash. Because the posterior predictive probability is similar for $SvM-c$ with three components and $SvM-p$ with three components, it is interesting to see these two aforementioned mean surfaces re-partition into three. As hinted by the lower half of the non-spiral surface, there is one component at either $\frac{\pi}{3}$ or $\frac{\pi}{4}$. This suggests a medium to semi-strong pull toward the second income category. Then, for the upper half of the non-spiral surface, $SvM-p$ takes two components. Interestingly enough, as seen in Figure 3.6e and 3.6f, the areas of high probability for $SvM-p$’s two components match the third component of $SvM-c$ with two components. These two mean components are centered at roughly $\frac{4\pi}{3}$ and $\frac{5\pi}{3}$, which represent semi-strong pushes away from the second and third income categories. In other words, the data already suggests the potential recession that might happen in the next year. Finally, the spiral mean surface changes as well and adopts characteristics of the other mean surface for the third mean surface for $SvM-c-3$ in 2007. As seen in 3.6f, it matches the first surface for neighborhoods with the lowest income proportions before rising rapidly to meet the other mean surface for neighborhoods with the second highest income proportions. It then comes back down to $\pi$, i.e. a pull toward the second income category, for neighborhoods with the highest income proportion.

### 3.6 Conclusion and future directions

In this chapter, we have introduced two types of hierarchical models to draw inference about the random directions for simplex-valued measurements and discuss how these models might be utilized. These approaches creatively use the data’s location within the simplex to do so. Because the average random direction across all location or the "mean surface" and the mixing probability are important, the "spatial" information is naturally fed into them. Indeed, $SvM-c$ is the model when this information is fed into the former whereas $SvM-p$ is the model when this information is fed into the latter. Then, to better understand how to set the hyperparameters of these models,
we derived the models’ prior circular means, variances, and correlations. We discussed how to
fit these models using sampling schemes that respect the geometry of the parameters of interest.
For instance, we used elliptical slice sampling to sample for the mean surface in $\mathcal{SvM}$-c. Using
these findings, we then applied the models to simulated data and then analyzed a data set of income
proportions and random directions observed for a set of census tracts in LA County from 1990 to
2010. There is evidence to suggest that the random direction is associated with the census tract’s
income proportions observed in a given year and not the tract’s physical location. This means that
the data set is a highly relevant real world example to apply our models to. Consequently, it is
noteworthy that the patterns our models discovered matches and potentially clarifies real world
economic trends during the same time period.
CHAPTER 4

Tree-based Mixture of Probability Simplices

4.1 Introduction

A natural way to describe how a collection of observations is generated is using admixture models. In these types of models, each observation from a collection has some collection specific probability of being generated from a particular distribution. A popular example of this framework is Latent Dirichlet Analysis (LDA) [12]. Here, the observations are words and the collection are documents. Each word has some probability of being generated according to a particular topic, which is a distribution on words. This topic probability varies for different documents.

Complication arises when there is further heterogeneity in the data. For instance, suppose that there are documents from multiple subjects. While this may be resolved by adding more distributions, a more compact way to deal with this issue is by organizing these distributions into a tree. In other words, each node corresponds to a distribution. Then, an observation can be generated with some probability from a distribution associated with a node on a path of the tree. This seems sensible because more common distributions are shared across various collections and can be put higher on the tree in order to be on more paths. Meanwhile, more unique distributions can be placed lower on the tree and thus on a few paths. Indeed, this inspired the development of the hierarchical LDA model with the nested Chinese Restaurant Process and the nested Hierarchical Dirichlet Process as priors [11, 54]. At a high level, these processes serve as priors for trees of any size and width. Further, the nested Hierarchical Dirichlet Process attempts to improve on the nested Chinese Restaurant Process by allowing a collection of nodes from the entire tree to serve as "paths".

Such models mentioned in the previous paragraph might also be called a tree-based mixture of admixtures because each path corresponds to an admixture model and the probability of choosing a path is akin to a mixing probability. If the distributions are discrete and the admixture model corresponding to each path can be represented as a simplex in some higher dimensional simplex, then the model can be denoted as a tree-based mixture of probability simplices. Because such
models are intuitive and potentially useful, it is desirable that the distribution induced by the model and the arrangement of the distributions on the tree are identifiable. In other words, it is useful to establish posterior consistency in the overall distribution, the component distributions themselves, and the arrangement of the distributions.

Schwartz provided a framework to establish posterior consistency of the overall distribution [64]. If it can be shown that the prior puts positive mass in all Kullback-Lieber neighborhoods of the true density, then the posterior weakly converges to that density. In the recent decades, this work was expanded upon and applied to more complicated models. For instance, Ghosal et al. and Barron et. al. built on this work to provide conditions for strong consistency for Dirichlet processes and Bayesian nonparametric models [10, 27]. This popularized the use of these techniques in the field of Bayesian nonparametrics. Other areas for which posterior consistency was demonstrated include time series [17], Gaussian processes priors for binary regression [28], and semi-parametric regression models [8]. Another key work that extended the framework for posterior consistency of the distribution by providing alternative conditions is a paper by Walker [69]. As a final example, Nguyen was able to demonstrate posterior consistency for the LDA model even though it was not the primary focus of the paper [53]. As a result, while it has not been done, it seems possible to extend these ideas in the discrete case.

However, even if we also examine work on posterior convergence rates because posterior convergence necessarily implies posterior consistency, there is very little theoretical work on posterior consistency of the distributions themselves and their tree-based organization. While there has been some work on identifying the posterior convergence rate of mixture models in both univariate and multivariate settings [15, 29, 30, 33, 52], the closest is still the aforementioned paper on the posterior consistency for the LDA model. This makes sense because such models are challenging to think about. In addition to the tree-based mixture being a mixture, the admixture model is already an extension of the mixture model because each observation follows a different proportion of the mixing distributions. This complicates how well an individual distribution can be understood as different distributions are revealed in a disimilar manner by each observation. The tree arrangement of the distribution adds further complexity because in addition to needing to learn the arrangement, certain distributions, such as the ones at the bottom, might not be observed as much as distributions on top of the tree [54].

Still, it is worthwhile to understand these models theoretically. Practitioners might wonder if it is possible to identify the distribution or the tree despite their complexity. Such information might give further confidence in the significance of the estimated structure. Further, the development of theory might lead to better methodology. For instance, while it is a simpler model, ideas from Nguyen’s paper on the contraction of polytopes were used to develop geometric algorithms [53, 74, 75, 76]. As a result, we begin to tackle this problem by focusing on the case in which the depth and width of
the tree are fixed. We also assume that the distribution in the admixture model is discrete and that the admixture itself can be represented as a probability simplex. In other words, we are interested in the tree-based mixture of probability simplices case.

We are particularly interested in a deeper theoretical understanding of these models in the finite case through the geometric perspective. First, after going over some relevant background, we extend previous work on the consistency of probability polytopes to establish posterior consistency of the mixtures’ distribution [53]. Next, we discuss a few assumptions in order to establish identifiability of the tree structure in the noiseless case. The argument can also be used to demonstrate posterior consistency of the tree-based mixture of admixture models with additional assumptions. As part of this discussion, two new metrics, a path metric and a tree metric, are introduced. The first metric is so named because it compares trees by summing the Hausdorff metric between simplices defined by the trees’ path. The other metric uses the $L_2$ distance between the vertices themselves and weights the difference according to the vertices’ position on the tree. By construction, the tree metric is more fine than the path metric.

We leverage these insights to develop alternative ways to fit these tree-based mixtures and to compute these metrics. In particular, we leverage the geometric insights to develop a fast way to fit tree-based mixture of vertices. Based on the path metric, this approach aims to minimize what we call the projected simplicial distance while controlling for the volume. The projected simplicial distance is the $L_2$ distance from the mean proportion of words observed in the document to the closest simplex. Because it builds on earlier work that can estimate the vertices of a single simplex, it also involves merging vertices in a bottom-up approach that preserves the tree structure. We also describe how these geometric insights might be applied to a Gibbs sampler. Finally, greedy approaches are used to compute the path and tree metric for balanced trees. These approaches use recursion and start from the root node.

Finally, we evaluate our methods by conducting a simulation study and applying our methods to a subset of New York Times articles from 2000. For the former, we explored the effect of increasing the dimension and the size of the tree while holding one of these aspects constant. To evaluate our performance, we compared the estimates using the path-based metric, tree-based metric, and the projected simplicial distance. In addition, to compare against Latent Dirichlet Allocation, we minimized the sum of the $L_2$ distance between the closest vertices. Then, because simulation demonstrates that the geometric methods are fast and have good performances, we use these methods and Variational Bayes to explore the New York Times data set. In addition to the estimated topics being meaningful, interpreting this tree reveals groups of articles related to events from 2000, such as the Florida recount during the 2000 United States presidential election. It also uncovers interesting semantic relationships.

The rest of the chapter is organized as following. First, we go over some necessary background.
After going over some notation, we introduce the path-based and tree-based metric. We then discuss when there is posterior consistency in the distribution and structure of the model. Next, we lay out a more specific version of the model and discuss how to fit using a Gibbs sampler, Variational Bayes, and geometric approaches. Because it also is a computation challenge, we show how to compute the metric. Then, we applied these methods to simulation and a subset of New York Times articles from 2000. We end with a short conclusion.

4.2 Background

Because our chapter aims to explore tree-based models from a geometric perspective, this background section will clarify what we mean by this viewpoint and describe the theoretical results and algorithms that can be derived from this viewpoint. We will focus on the discrete admixture or LDA case. Here, the object of interest are the topics, \( \theta_k \in \Delta^D \) for \( k = 1, \ldots, K \) and \( K, D \in \mathbb{N} \). Once the topics are known, the proportions of words observed in a document for the LDA model is simply a convex combination of these topics with some noise, i.e. \( \sum_k \beta_k \theta_k + \epsilon \) for some \( \beta_k \in \Delta^K, \epsilon \in \mathbb{R}^D \). In other words, the topics define some polytope, \( G \), from which the generating discrete probability distribution for each document is drawn from. As a result, to study the convergence rate of these models, it is sufficient to understand the convergence rate of the topics or the polytope they define.

There are two metrics that can be used. One is the minimum-matching distance:

\[
 d_M(G, G') = \max_{\theta \in \text{extr}(G)} \min_{\theta' \in \text{extr}(G')} \| \theta - \theta' \| \lor \max_{\theta' \in \text{extr}(G')} \min_{\theta \in \text{extr}(G)} \| \theta - \theta' \|.
\]

Notice that in this notation, the topics are the extreme points of the polytope. Meanwhile, an alternative metric is the Hausdorff metric:

\[
 d_H(G, G') = \max_{\theta \in G} d(\theta, G') \lor \max_{\theta' \in G'} d(\theta', G).
\]

Here, \( d(\theta, G') \) is defined to be \( \inf\{\|\theta - \theta'\|, \theta' \in G'\} \). While \( d_H(G, G') \leq d_M(G, G') \) by construction, it can be shown that \( d_M(G, G') \leq C_0 d_H(G, G') \) for some positive constant \( C_0 \) under some mild assumptions [53].

Then, the following results can be established [53]. First, under some geometric assumptions and some mild assumptions, the convergence rate can be established to be:

\[
 \left( \frac{\log(m)}{m} + \frac{\log(n)}{n} + \frac{\log(n)}{m} \right)^{\frac{1}{\alpha + \beta}}.
\]

Note that \( m \) represents the number of documents, \( n \) represents the number of words in the document,
\( p = (K - 1) \wedge D \), and \( \alpha \) denotes a regularity level near the boundary of the polytope defined by the topics. If the number of topics are also known or all topics are also at least a certain distance apart, then the convergence rate can be improved to be:

\[
\left( \frac{\log(m)}{m} + \frac{\log(n)}{n} + \frac{\log(n)}{m} \right)^{\frac{1}{2(1+\alpha)}}.
\]

As part of establishing this convergence rate, a Kullback-Leibler upper bound can be established on the induced distribution. In particular, it can be shown that if \( S[n] \) denotes a sample, \( p_{S[n]|G} \) the probability of observing a sample given a polytope, and \( p_{\eta|G} \) the probability of a point within the polytope, 

\[
K(p_{S[n]|G}, p_{S[n]|G'}) \leq \frac{n}{c_0} W_1(p_{\eta|G}, p_{\eta|G'}). 
\]

Algorithms were developed to take advantage of this geometric perspective and fit the discrete admixture model to data. At a high level, while it is challenging to identify the extreme points of a polytope by searching the space, the interior of a polytope can be used to identify its extreme points. In particular, if the interior is clustered correctly, then the extreme points lie on a ray from the center of the polytope in the direction of the clusters’ centers. The first such algorithm to apply this approach is Geometric Dirichlet Means or GDM [76]. Using the proportions of words observed in each document lie within the polytope, the algorithm applies k-means to cluster these proportions. The estimated extreme points are the points that are the intersections between the rays extending from the center of all proportions to the center of the clusters and the boundary. However, this approach may not work if the polytope is not an equilateral or the proportions do not lie near the topics.

One improvement to GDM is Voronoi Latent Admixture or VLAD [75]. Now, the clusters of interest are Centroidal Voronoi Tessellations. To define these tessellations, we first define a Voronoi cell. Given an open set, \( \Omega \), a distance function, \( d \), a probability density function, \( \rho \), a collection of points, \( c_1, c_2, \ldots, c_K \), a Voronoi cell for \( c_k \) is the set

\[
V_k = \{ x \in \Omega : d(x, c_k) < d(x, c_l) \text{ for any } l \neq k \}.
\]

These collections of cell are a Centroidal Voronoi Tessellations if \( \rho_k(A) = \frac{\rho(A \cap V_k)}{\rho(V_k)} \) and \( c_k = E_{\rho_k}(x) \) for any set \( A \) and \( k = 1, 2, \ldots, K \). In other word, \( c_k \) is a centroid for \( V_k \). These tessellations are of interest because in the scenario that the proportions of words observed in documents are the convex combination of the topics, i.e. the noiseless case, \( c_1, c_2, \ldots, C_K \) must lie on the line segments connecting the overall center of the polytope to the topics. These centroids are also easy to compute as they can be discovered by performing k-means on the left singular vectors of the centered data. Meanwhile, the extension parameter can be found via a Monte Carlo technique on a symmetric Dirichlet distribution with a known or estimated parameter. Under certain assumptions, VLAD can
be shown to be a consistent estimator in the noiseless case and for the LDA model. Indeed, the rate of convergence can be established.

Both VLAD and GDM require the number of topics to be known. If the number of topics are not known, the Conic Scan and Cover (CoSAC) algorithm can be used [74]. After centering the data and making all points "active", the algorithm identifies topics in the following manner. Given an "active" set of points, the data point farthest away according to the $L_2$ distance is identified as a topic. Then, all data points within a certain cosine distance of that point are removed from the set of "active" points. In other words, these points are "covered" by a cone emanating from the topic. The algorithm then repeats this process until there are no more "active" points or the remaining points within the "active" set are too close to the center. If the observed proportions are the convex combination of topics with noise, there are further adjustments that one can make to the estimated topics to account for the noise. While there is no theory for CoSAC, it can still be understood in the same framework as VLAD and GDM. After all, while the steps might be reversed, CoSAC creates a procedure to identify the clusters when the number of clusters are unknown and extends in the direction of the clusters to discover the topic.

4.3 Model

4.3.1 Notation and model setting

Now, we restate our notation for clarity as we describe the setting for our models. We assume that the vertices represented by $\theta$ can be organized in a finite tree structure with each node on the tree corresponding to a vertex. Here, $\theta \in \Delta^{D-1}$ for some $D \in \mathbb{N}$ and the vertices are assumed to be distinct. The tree is assumed to have a depth of $J$ and $K$ nodes in total. We index these nodes from $1, 2, \ldots, K$ by level and then from left to right at a level. More formally, if there exists nodes, $V_1, V_2, \ldots, V_K$, and a vertex associated with each node, $\theta_k$, for $k = 1, 2, \ldots, K$, let the
We then assign probability to these simplices in the same manner as described in [53]. In other words, we can define a random vector \( \eta \) words, we can define a random vector

\[
X = \{\{V_k\}_{k=1}^K, E, \{\theta\}_{k=1}^K, V_1\}.
\]

As discussed previously, the vertices associated with each node and the arrangement of these vertices are of interest to us. Then, a path is a collection of node indices from the root to the leaf such that there is only one node indexed at each level and except for the node at the first level, each node has to be one of the children nodes of the one at the level above. Let \( \varphi \) denote a path ordered by level such that \( \varphi_i \) is the node index at level \( i \). Further, denote a subset of paths as \( \phi(T) \) and the set of all paths as \( \Phi(T) \). Let \( I_\rho \) indicate the number of paths, i.e. \( |\Phi(T)| \).

We also want to give some probability to the levels on the path and the probability of picking a path. To that end, we use \( \beta \in \Delta^{J-1} \) and \( \pi \in \Delta^{I_\rho-1} \) respectively. We assume that \( \pi_i > 0 \) for all \( i = 1, 2, \ldots, I_\rho \).

It is possible to define a probability simplex based on the vertices indexed by these paths or for simplicity, defined by the paths. First, to define a simplex, let \( \text{conv}(\cdot) \) denotes the convex hull of a set of points. Then, set \( G_i \) to be \( \text{conv}(\{\theta_k\}_{k \in \varphi_i}) \) for \( i = 1, 2, \ldots, I_\rho \). Because the vertices lie in a simplex, these convex hulls are non-degenerate simplices. As a result, the number of vertices cannot exceed one plus the ambient dimension so the ambient dimension of the simplices is \( J - 1 \).

We then assign probability to these simplices in the same manner as described in [53]. In other words, we can define a random vector \( \eta_i = \sum_j \beta_j \theta_{\varphi_i} \). If we let \( \beta \) be distributed according to some random distribution \( P_{\beta|\gamma} \) for some parameter \( \alpha \), such as a Dirichlet distribution, this induces a random distribution with support, \( G_i \). We denote this distribution as \( p_{\eta|G_i} \) and leave further details of this derivation to [53].

Before we define the tree-based mixture of probability simplices, we need to discuss the data. Let there be \( M \) collections with \( N_m \) observation for collection \( m, m = 1, 2, \ldots, M \). If the sample size is the same, we drop the subscript. In other words, a particular collection is indicated as \( S^m_{[n]} = (X_{m,n})_{j=1}^{N_m} \). Then, this collection is assigned some path indicated by \( z_m, z_m \in 1, 2, \ldots, I_\rho \), for \( m = 1, 2, \ldots, M \). The probability simplex defined by the path of \( z_m \) is used to generate a random vector, \( \eta_m \). Here, \( \eta_m \) is distributed according to \( p_{\eta|G,z_m} \) from the previous paragraph. Then, \( S^m_{[n]} \) is assumed to be a draw from a multinomial distribution, \( \text{Multi}(\eta_m) \). More specifically, we have that \( X_{m,n} \in \{1, 2, \ldots, D\} \) such that \( p(X_{m,n} = d \mid \eta_m) = \eta_{m,d} \).

Combining this with our earlier discussion on the probability simplex, we have the following distributions. First, for an implied \( z_m = i \), we have that

\[
p_{\eta \times S_{[n]}|G_i}(\eta_m, S^m_{[n]}) = p_{\eta|G_i}(\eta_m) \times \prod \eta_{m,d}^{\lfloor X_{m,n} = d \rfloor}. \quad (4.1)
\]

Then, we can marginalize out \( \eta_m \) to get a distribution for \( S_{[n]} \):

\[
p_{S_{[n]}|G_i}(S^m_{[n]}) = \int_{G_i} \prod \eta_{m,d}^{\lfloor X_{m,n} = d \rfloor} \, dp_{\eta|G_i}(\eta_m). \quad (4.2)
\]
We denote this as $p_{S_{[n]}|G_i}$. If we then additionally integrate out $z_m$, we can get a tree-based mixture distribution for $S_{[n]}$:

$$p_{S_{[n]}|T}(S_{[n]}^m) = \sum \pi_i p_{S_{[n]}|G_i}(S_{[n]}^m).$$  \hspace{1cm} (4.3)

This distribution will be indicated as $p_{S_{[n]}|T}$.

Another distribution of interest is the noiseless case. Here, given collection $S_{[n]}^m$, $\frac{1}{n} \sum_n I(X_{m,n} = d) = \eta_{m,d}$ for $d = 1, 2, \ldots, D$. In other words, $\frac{1}{n} S_{[n]}^m = \eta_m$. Then, the noiseless version of the tree-based mixture distribution is the following:

$$\tilde{p}_{S_{[n]}|T}(S_{[n]}^m) = \sum \pi_i p_{\eta|G_i} (\eta_{m,n}) I(\eta_{m,n} \in G_i).$$  \hspace{1cm} (4.4)

As we demonstrate later in the chapter, the noiseless case is useful because it is a simpler distribution that can be analyzed and still provide valuable insights for understanding $p_{S_{[n]}|T}$.

We end this section with a few notational notes. As hinted above, we bold a variable if we are interested in the entire collection. If we are interested in all but the current index, we will attach a negative index to the bolded variable. For instance, $\theta$ denotes the collection of all vertices and $\theta_{-k}$ denote the collection of all vertices without $\beta_k$. As another example, let $\beta$ denote the collection of all $\beta$’s and $\beta_{-m}$ the collection all $\theta$’s without $\theta_m$. Let $P$ denote the set of paths. Finally, we will be explicit when we are referring to the noiseless case. If no indications are given or the typical case is mentioned, $S_{[n]}$ is assumed to be distributed according to a multinomial distribution and the distributions related to that are being used.

### 4.3.2 Metrics on the tree-based mixtures

For theoretical reasons and to evaluate how well our models fit the data, we will define two metrics on the vertices associated with nodes arranged in a tree: a path-based metric and a tree-based metric. The first allows us to extend the work of [53]; the second is a more intrinsic definition based on the organization of vertices imposed by the tree. For the trees, $T$ and $T'$, that we compare, we make the following assumptions:

(A1) $T$ and $T'$ have depth $J$;

(A2) $I_p = I_{p'}$;

(A3) the number of nodes with a certain number of children are the same at each level;

(A4) each node except the children node has at least two children node.
Then, before we define our metrics, we need to define the permutations that these metrics will use. Note that we will abuse notation and use the same notation for the permutation and the function on paths induced by the permutation.

For our path-based metric, we need to first define a path-preserving permutation.

**Definition 4.3.1.** Given two trees \( T \) and \( T' \) that satisfy assumptions (A1) – (A3), let \( \Phi(T) \) and \( \Phi(T') \) be two sets of paths. Then, let \( \sigma \) be a permutation from \( \Phi(T) \) to \( \Phi(T') \). Suppose that for some \( i = 1, 2, \ldots, I \rho \) and \( j = 1, 2, \ldots, J \rho \), \( \sigma(\varphi_i) = \varphi'_{j'} \). Further, suppose that for all \( j' \leq j \), \( j' = 1, 2, \ldots, J \rho \), \( \sigma(\varphi_i) = \varphi'_{j'} \). If this is true for any \( i = 1, 2, \ldots, I \rho \) and \( j = 1, 2, \ldots, J \rho \), call \( \sigma \) a path-preserving permutation and denote it as \( \sigma_p \).

We can use this permutation to define a metric, which we show to be a valid metric in the appendix.

**Definition 4.3.2.** Suppose that there are two trees, \( T \) and \( T' \), that satisfy assumptions (A1) – (A4). For \( i = 1, 2, \ldots, I \rho \), set \( G_{\varphi_i} \) to be the convex hull of vertices on the nodes along the path \( \varphi_i \) from tree \( T \) and \( G'_{\varphi_i} \) to be a similar convex hull for topics from tree \( T' \). Let \( \sigma_p(\varphi_i) \) indicate the path mapped to by the permutation defined in Definition 4.3.1. If \( d_H(G_{\varphi_i}, G'_{\sigma_p(\varphi_i)}) \) is the Hausdorff metric, define the path-based metric as following:

\[
d_p(T, T') := \min_{\sigma_p} \sum_i d_H(G_{\varphi_i}, G'_{\sigma_p(\varphi_i)}).
\]

We again need to first define a permutation for our tree-based metric.

**Definition 4.3.3.** Let \( T \) be a tree that satisfies assumptions (A1) – (A3). For \( j = 1, 2, \ldots, J \), define \( \sigma_j^T \) to be a permutation of the nodes of \( T \) at level \( j \) and all nodes below such that the parents and children of all nodes of \( T \) are preserved. Call this permutation a sub-tree preserving permutation.

For our tree-based metric, we define it recursively in the following manner. We will abuse our notation and use \( \theta_k \) to indicate the vertex that corresponds to \( V_k \).

**Definition 4.3.4.** Suppose that \( T \) and \( T' \) are two trees that satisfy assumptions (A1) – (A3). At depth 1, if \( \beta_1 \) and \( \beta'_1 \) are the topics at the root node, set

\[
d_1(T, T', \sigma_1^T) := \|\theta_1 - \theta'_1\|.
\]

At depth \( j \), let \( \sigma_j^T \) be the permutation defined in Definition 4.3.3. Further, let \( \sigma_j^T (\cdots (\sigma_2^T(V_k))) \) indicates which node \( V_k \) is mapped to after applying \( \sigma_j^T \) to the tree permuted by \( \sigma_2^T \) and so on until
\(\sigma_{j-1}^T\) in that order. Then, if \(\ell(\cdot)\) denotes the level of a vertex on the tree, set
\[
d_j(T, T', \sigma_j^T) := \sum_{k: \ell(V_k) = j} \left( \sum_i \mathbb{I}(k \in \varphi^i) \right) \left\| \theta_k - \theta'_{\sigma_j^T(\cdots(\sigma_2^T(V_k)))} \right\|.
\]

The tree-based metric is thus defined as following:
\[
d_T(T, T') := \min_{\sigma_2^T, \sigma_3^T, \ldots, \sigma_J^T} \sum_{j=1}^J d_j(T, T', \sigma_j^T). \tag{4.5}
\]

To understand the relationship between these two metrics, we first note that if there are duplicate topics, the tree-based metric can detect the differences whereas the path-based metric cannot. This is because the path-preserving permutation allows nodes that lie on different branches, but have the same vertices as parents to be swapped. As a result, the tree-based metric is more fine than the path-based metric. However, even under our assumption that there are distinct topics, this will still hold. Indeed, this is by design and so we leave the proof to the appendix. However, while we can demonstrate that the tree-based metric is greater than the path-based metric, it is less clear how to come up with mild assumptions like Property (A2) from [53] to prove the inverse. Fortunately, this is not necessary because we are not interested in contraction rates in this chapter.

### 4.3.3 Identifiability and consistency

With these metrics, we now work to show posterior consistency of the distribution and identifiability. We accomplish the former by establishing an upper Kullback-Leibler (KL) bound and the latter by making geometric arguments.

To derive this upper bound, we extend Lemma 6 from [53]. As we show in the following lemma, we are able to do so through a decomposition of the KL by path and a clever re-arrangement.

**Lemma 4.3.1.** Let \(T\) and \(T'\) be two trees that satisfy assumptions (A1) and (A2). Assume that for all polytopes defined by the paths in \(P_T\) and \(P_{T'}\) indexed by \(i, i' \in 1, 2, \ldots, \mathcal{I}_\rho, G_i, G_{i'} \in \Delta^D\) are closed convex subsets such that for any \(\eta \in G_i \cup G_{i'}, \min_d \eta_d > c_0\) for some constant \(c_0 > 0\) and \(\min_i \pi_i > c_1\) for some constant \(c_1 > 0\). Then, if \(\sigma\) denotes a permutation between \(\Phi(T)\) to \(\Phi(T')\),
\[
K(p_{s_{[n]|T}:p_{s_{[n]|T'}}} \leq \sum_i \pi_i \frac{N}{c_0} W_1(p_{\eta|G_i}, p_{\eta|G_{\sigma(i)}}) + \sum_{\mathcal{I}_\rho} \frac{1}{c_1} |\pi_i - \pi'_{\sigma(i)}|.
\]
Proof. Let $S_i = \{x_{n}\}_{n=1}^{N}$ and associate it with a vector $\eta_i$ for each path from $T$ and $\eta'_i$ for each path from $T'$ for $i = 1, 2, \ldots, I$. We have that

$$p_{S_i|T}(S_i) = \sum_i \int_{G_i} p(S_i) dP_{\eta_i}(\eta_i) \pi_i.$$\[12pt]

Further, because of assumption (A2), $\sigma$ is a bijective mapping. Then, assumption (A1) allows us to define a coupling $Q_z$ between $\eta_i$ and $\eta'_i$. It follows from the convexity of the KL divergence, Jensen’s inequality, and our assumptions that

$$K(p(S_i|T), p(S_i|T')) = K \left( \sum_i \int_{G_i} p(S_i) dQ_z(\eta_i, \eta'_i) \pi_i, \sum_i \int_{G_i} p(S_i) dQ_z(\eta_i, \eta'_i) \pi_i' \right) \leq \sum_i K \left( \int_{G_i} p(S_i) dQ_z(\eta_i, \eta'_i) \pi_i, \int_{G_i} p(S_i) dQ_z(\eta_i, \eta'_i) \pi_i' \right) \leq \sum_i \pi_i K(\int_{G_i} p(S_i) dQ_z(\eta_i, \eta'_i)) \leq \sum_i \frac{N}{c_0} W_1(p_{\eta_i|G_i}, p_{\eta'_i|G_i}) + \sum_{I'} \frac{1}{c_1} |\pi_i - \pi'_i|.$$

With Lemma 4.3.1, we can establish the following theorem on the neighborhood of prior support a tree, $T_0$, its vertices, $\theta_1^*, \theta_2^*, \ldots, \theta_K^*$, and some path probability, $\pi^*$. In particular, if $T^*$ is the entire space of trees with its vertices and path probability, define

$$B_K(T_0, \delta) = \{ T \in T^* \mid K(p_{S_i|T_0}, p_{S_i|T}) \leq \delta^2, K_2(p_{S_i|T_0}, p_{S_i|T}) \leq \delta^2 \},$$

where $K_2(p_{S_i|T}, p_{S_i|T_0}) := \int p_{S_i|T} \left( \log \frac{p_{S_i|T}}{p_{S_i|T_0}} \right)^2$. \[12pt]\]

Theorem 4.3.1. Suppose that the prior, $\Pi$, exists with the following assumptions:

- any tree satisfies assumptions (A1) and (A2),
- each of $\theta_1, \ldots, \theta_K$ is bounded away from the boundary of $\Delta^{D-1}$, i.e. $\min_d \theta_{k,d} > c_0$ for all $j = 1, 2, \ldots, K$. 

75
• \( \min_i \pi_i > c_1 \) for some constant \( c_1 > 0 \),
• \( \Pi(\|\theta_k - \theta_k^*\| \leq \epsilon \forall k = 1, 2, \ldots, K) \geq c'_0 \epsilon^{kd} \) for some \( c'_0 > 0 \) and any small \( \epsilon \),
• \( \Pi(\|\pi - \pi^*\| \leq \epsilon) \geq c''_0 \epsilon^{kd} \) for some \( c''_0 > 0 \) and any small \( \epsilon \), and
• \( \Pi(\|\theta_k - \theta_k^*\| \leq \epsilon) \) is independent of \( \Pi(\|\pi - \pi^*\| \leq \epsilon) \) for any small \( \epsilon \).

Then, for any \( T_0 \) in the support of \( \Pi \), any \( \delta > 0 \), and \( n > \log(1/\delta) \),

\[
\Pi(T \in B_K(T_0, \delta)) \geq c \left( \frac{\delta^2}{N^3} \right)^{2kd},
\]

where \( c \) is a constant that is a function of \( c_0, c_1, c'_0, c'_1 \), and \( I_\rho \).

Proof. As in the proof of Theorem 6 from [53], we need to show that \( \int \frac{p_{S[\pi]|T}}{p_{S[\pi]|T_0}} < M \) and

\[
h(p_{S[\pi]|T}, p_{S[\pi]|T_0}) \leq \epsilon < \epsilon_0 \text{ for some constant } M \text{ and } \epsilon_0. \]

We then have \( K_2(p_{S[\pi]|T}, p_{S[\pi]|T_0}) := \int p_{S[\pi]|T} \left( \log \left( \frac{p_{S[\pi]|T}}{p_{S[\pi]|T_0}} \right) \right)^2 = O(\epsilon^2 (\log M)^2) \), which we can use to establish the lower bound on our prior.

First, let \( G_i^* \) denote the simplex defined by path \( i \) for the tree \( T_0 \), its vertices, \( \theta_1^*, \theta_2^*, \ldots, \theta_K^* \), and some mixing probability, \( \pi^* \). If \( T \) is a random draw from \( T^* \) with vertices, \( \theta_1, \theta_2, \ldots, \theta_K \), and path probability, \( \pi \), let \( G_i \) denote the simplex defined by path \( i \) for the tree \( T \). Set \( \sigma \) to be a permutation between paths of \( T_0 \) and \( T \). This is possible due to Assumptions (A1) and (A2). Then, consider the event \( \|\theta_k^* - \theta_k\| \leq \epsilon \) for all \( k = 1, 2, \ldots, K \) and \( \|\pi^* - \pi\| \leq \epsilon \). Under this event, we can use a similar coupling as the one used in the proof of Theorem 6 from [53]. In other words, set the coupling distribution, \( Q_i \), for \( G_i \) and \( G_{\sigma(i)}^* \) so that for any \( (\eta_1, \eta_2) \), \( \eta_1 = \sum_j \beta_j \theta_{\varphi_j} \) and \( \eta_2 = \sum_j \beta_j \theta_{\varphi_j}^* \) for some \( \beta \in \Delta^{J-1} \). Hence, under this event, \( W_1(p_{\eta|G_i}, p_{\eta|G_{\sigma(i)}^*}) \leq \epsilon \). Then, based on Lemma 4.3.1 and Cauchy-Schwarz,

\[
h^2(p_{S[\pi]|T}, p_{S[\pi]|T_0}) \leq \frac{K(p_{S[\pi]|T}, p_{S[\pi]|T_0})}{2} \leq \frac{1}{2} \left( \sum_i \pi_i N \frac{W_1(p_{\eta|G_i}, p_{\eta|G_{\sigma(i)}^*})}{c_0} + \sum_k \frac{1}{c_1} |\pi_i - \pi_{\sigma(i)}^*| \right) \leq \epsilon \left( \frac{N}{2c_0} + \frac{\sqrt{I_\rho}}{c_1} \right),
\]

Meanwhile, we have that

\[
\frac{p_{S[\pi]|T}}{p_{S[\pi]|T_0}} = \sum_i \pi_i p_{S[\pi]|G_i} \leq \sum_i \sum_{\varphi_i} \pi_{\varphi_i} p_{S[\pi]|G_{\varphi_i}^*} \leq \frac{I_\rho}{c_1} \left( \frac{1}{c_0} \right)^N.
\]
As a result, \[ \sum_{S[n]} \left( \frac{p_{S[n]}|T}{p_{S[n]}|T_0} \right)^2 \leq \frac{\sqrt{T^2}}{c_1} \left( \frac{1}{c_0} \right)^N. \]

So, based on the first paragraph of this proof, \( K_2(p_{S[n]}|T, p_{S[n]}|T_0) \) must be the following order:

\[
O \left( \epsilon \left( \frac{N}{2c_0} + \frac{\sqrt{T^2}}{c_1} \right) \left( N \log \left( \frac{1}{c_0} \right) + \log \left( \frac{T^2}{c_1} \right) - \frac{1}{2} \log \left( \epsilon \left( \frac{N}{2c_0} + \frac{\sqrt{T^2}}{c_1} \right) \right) \right)^2 \right).
\]

Hence, if we set \( \epsilon \) to be \( \frac{\delta^2}{N^3} \), then \( K_2(p_{S[n]}|T, p_{S[n]}|T_0) \) is bounded by \( O(\delta^2) \) as long as \( N > \log(1/\delta) \). We get the result in the theorem by using independence and the lower probability bounds on the prior.

We now make some comments on this theorem. First, this theorem demonstrates that with a certain prior, there is enough probability in the KL ball around any tree, its vertices, and its path probability in the prior’s support. It is often realized in practice because the vertices and the path probability usually are both distributed according to Dirichlet distributions. These distributions are usually independent of each other and may be different. Next, according to [53], the probability for a KL ball around a single polytope is of the order \( \delta^2 N^3 k^d \). It is interesting that adding more simplices and a mixing probability only squares this probability. However, this might be a reflection of the prior assumptions. Before the probability of the simplex comes in, the model has to first choose a simplex. Due to independence, this extra "layer" to the model needs to have the same prior probability as the prior probability for the simplex. A different prior might be desired to reflect the fact that under certain conditions, once the vertices are known, the tree and mixing probability can be determined as well. This class of priors might give more probability to the KL ball than the current class. However, we leave discussion of this to future work because we already demonstrated that there is a class of priors for which the probability of a KL ball is lower bounded and as discussed previously, this class of prior is commonly used.

Next, we want to prove consistency for the structure. Formally, this is specified as showing that \( V(p_T, p_{T'}) \to 0 \implies d_P(T, T') \to 0 \). At a high level, we particularly want to show that the vertices converge to the "true" vertices and their arrangement correspond to the "true" tree. To start, we begin with the simpler problem of identifiability in the noiseless case. This involves making assumptions so that we cannot re-partition the space enclosed by the union of these polytopes. First, we introduce assumptions on the probability measure:

(A5) \( p_{\eta|G_i} \) is absolutely continuous;

(A6) given \( A \subset G_i, p_{\eta|G_i}(\eta \in A) > 0 \);

(A7) if \( p_{G_1}, p_{G_2}, \ldots, p_{G_{\mathcal{I}}} \) belong to a family of distributions, then \( \bigcup_{i \in \mathcal{I}} \pi_i p_{G_i} \) do not belong to the same family of distribution for any subset of path indices, \( \mathcal{I} \).
In other words, we assume the probability measure is absolutely continuous, non-zero, and distinct. We also need to make a geometric assumption on these simplices. To do so, we must first define collinear.

**Definition 4.3.5.** Vertices $\theta_{k_1}, \theta_{k_2}, \theta_{k_3}$ are **collinear** if there exists some $t \in \mathbb{R}$ such that $\theta_{k_1} = (1-t)\theta_{k_2} + t\theta_{k_3}$.

Based on this definition, our assumption is the following.

(A8) No vertex is collinear with any other two vertex.

This last assumption gives us the following crucial lemma, which we prove in the appendix.

**Lemma 4.3.2.** Suppose that we have a tree, $T$, such that all paths have length $J > 1$. Assume that (A8) holds. Then, the union of simplices defined by the tree have at least two exposed vertices in the noiseless case.

As shown below, we can use this lemma to establish identifiability because it means that there will always be two exposed vertices.

**Theorem 4.3.2.** Suppose that we have trees, $T$ and $T'$, such that (A1) and (A5) – (A8) holds and each node except the leaf nodes has at least two children. Then, if $V(\bar{p}_{S[n]|T}, \bar{p}_{S[n]|T'}) = 0$ in the noiseless case, $d_P(T, T') = 0$.

**Proof.** Let the tree of depth $J$ be indicated as $T$. Note that based on our previous lemma, we have at least two exposed nodes. Due to distinct probabilities from Assumption (A7), we can find rays emanating from these exposed extreme points. Due to the collinear constraint, assumption (A8), if the two nodes have rays pointing to each other, then we know that the two exposed nodes belong to the same simplex. The collinear constraint also allows us to assume that the intersections from the remaining rays identifies that simplex if the two nodes belong to the same simplex. On the other hand, if the two nodes don’t have rays pointing to each other, we can examine their intersections. Because of the distinct probability assumption, collinear constraint, and assumptions (A5) and (A6), we can identify which intersections are extreme points. As the simplices must at least share the root, we again have identified two nodes from the same simplex and can use the collinear constraint to identify that simplex. If we consider the tree without the identified simplex, we can again apply the lemma because of our assumptions and use the two nodes to identify another simplex. As a result, by repeating this, we can identify all simplices defined by the tree.

Now that we have discovered all simplices, we can identify the tree. First, we examine all simplices that share $J - 1$ nodes. Under our assumption that each node except the leaf node having at least two children, each simplex must share $J - 1$ node with at least one other simplex.
The unmatched nodes are the leaf nodes. Let these matched simplices be denoted as subtrees, \( \tilde{T}_1^1, \tilde{T}_2^1, \ldots, \tilde{T}_j^1 \) for some \( J_1 \). We can then merge these subtrees based on the subtrees that share \( J - 2 \) nodes. Again, each subtree will be merged with at least one other subtree. Any unshared nodes must be the nodes at level \( J - 1 \). By repeating this process, we can identify the nodes at each level. Because each node has at least two children, the level for each node is uniquely identified. We can also examine how the subtrees are merged to understand the tree structure. As a result, under these assumptions, there is only one tree-based mixture of vertices that describes the data in the noiseless case.

In addition to establishing identifiability in the noiseless case, we are interested in establishing posterior consistency of the structure in the multinomial noise case. Existing techniques focus on isolating an \( \epsilon \)-ball within one simplex based on the Hausdorff metric. Using the path-based metric or a metric based on the maximum Hausdorff distance might serve as a natural extension for the tree-based mixture of simplices. However, these metrics can be challenging to use because there might be another simplex that intersects with the \( \epsilon \)-ball based on these metrics and there may be no other location to do the comparison. To address this concern, we might want to use the minimum non-zero Hausdorff metric between two simplices. Now, when a new pair of simplex are introduced, the minimum Hausdorff might remain unchanged or become smaller. It might then be possible to find an \( \epsilon \)-ball in either case. Unfortunately, such a metric is too weak. As a trivial example, suppose the Hausdorff metrics between two sets of two simplices are 1 and \( \frac{1}{M} \), \( M \) the number of observations. Then, the minimum non-zero Hausdorff metric goes to zero as \( M \) goes to infinity, but the Hausdorff metric between one pair of simplices is still 1.

As a result, we consider the following distance instead:

**Definition 4.3.6.** Suppose that we have two trees, \( T \) and \( T' \), such that assumption (A1) hold. Define the union Hausdorff distance to be the following: \( d_{UH}(T, T') := d_H(\bigcup G_i, \bigcup G'_i) \).

The union Hausdorff distance is not a metric because consider the following trivial case. Suppose that we have two simplices, \( G_1 \) and \( G_2 \), such that \( G_2 \subset G_1 \) and \( G_1, G_2 \) have all but one vertices in common. Then, let this hold for another pair of simplices, \( G'_1 \) and \( G'_2 \). If \( G_1 = G'_1 \), but \( G_2 \neq G'_2 \), then the union Hausdorff distance between the pairs of simplices is zero, but the pairs are not the same. However, we can establish the following with this distance.

**Theorem 4.3.3.** Suppose that we have two trees, \( T \) and \( T' \), such that assumptions (A1) and (A2) hold. Suppose also that \( \min_i \pi_i > c_1 \) for some constant \( c_1 > 0 \). Assume that for all \( i \in 1, 2, \ldots, I_p \), as defined in [53], both \( p_{\eta|G_i} \) and \( p_{\eta|G'_i} \) and any restrictions are \( \alpha - regular \) densities on \( G, G' \). Further, \( G, G'_i \) have the thick body property discussed in [53]. Then, if \( V(p_{S[n]|T}, p_{S[n]|T'}) \to 0 \), \( d_{UH}(T, T') \to 0 \).
At a high level, the proof seeks to establish an "inverse bound" on the total variation by following the proof of Theorem 5 from [53]. Let $\hat{\eta}$ represent the observed proportions and $\eta_i$ be some random vector from the simplex defined by path $i$. We can then demonstrate that $p_{S[n]|T}(|\hat{\eta} - \eta_i| > \epsilon)$ is lower bounded in a similar manner as before. Meanwhile, the parts of the simplices defined by paths from one tree that do intersect with any of the simplices defined by paths from the other tree can be re-partitioned and described as the union of polytopes. Then, the union Hausdorff distance gives us enough control that we can find an appropriate $\epsilon$-ball based on the distance in one of these polytopes. This gives us a lower bound on $p_{S[n]|T}(|\hat{\eta} - \eta_i| \leq \epsilon)$. However, because of the proof’s length, we defer it to the appendix.

To show how this theorem can be applied, we need to discuss when and how the union Hausdorff distance and the path-based metric are related. Based on our earlier discussion, we obviously need the following assumption.

(A9) $\text{int}(G_i) \cap \text{int}(G_{i'}) = \emptyset$ and $\text{int}(G_{i'}) \cap \text{int}(G_i) = \emptyset$ if $i \neq i', i, i' \in 1, 2, \ldots, I_{\rho}$.

This assumption and a few other assumptions allow us to prove the following corollary, which demonstrates the association between the union Hausdorff distance and the path-based metric.

**Corollary 4.3.1.** Suppose that we have two trees, $T$ and $T'$ such that Assumptions (A1) – (A4) and (A9) hold. For the simplices defined by the paths of the trees, assume that any two simplex defined by a path that only differ at the leaf nodes are affine independent. If $d_{UH}(T, T') = 0$, then $d_{T}(T, T') = 0$.

**Proof.** Because we assume that the interiors of the simplices defined by the path are distinct, no simplex can be contained in another simplex. Further, there must be no common face between simplices defined by paths that do not only differ by the leaf. As a result, the simplices that make up the union of simplices defined by these paths must be unique. Meanwhile, suppose that $G_1$ and $G_2$ are simplices defined by paths that differ only by the leaf. Then, because of the affine independence assumption and assumption (A1), the dimension of the union $G_1$ and $G_2$ must be $J$. As the dimension of $G_1$ and $G_2$ must be $J - 1$, the simplices that make up the union of simplices defined by these paths must again be unique. As a result $d_{UH}(T, T') = 0$ means that $d_{H}(G_i, G'_{\sigma(i)}) = 0$ for some permutation, $\sigma$. This permutation is bijective according to Assumption (A2). We can again use the argument to establish identifiability in the noiseless case once the simplices have been identified because of assumption (A4). As a result, this permutation can be made into a path-preserving permutation. \qed

As a result, while it may be difficult to show that as $V(p_{S[n]|T}, p_{S[n]|T'}) \to 0$, $d_{P}(T, T') \to 0$, the theorem and corollary can be used to show that if $d_{UH}(T, T') \to 0$, $d_{P}(T, T') = 0$ when $d_{UH}(T, T') = 0$. In other words, we can establish posterior consistency in the "rough" sense.
Because the argument in the corollary can be used to establish that the tree-based metric will also be zero, there is a similar type of consistency for the tree-based metric.

4.4 Model fitting and computation

We are now interested in fitting this tree-based mixture of probability simplices. However, we must first be more specific about the model we are interested in fitting. After doing so, we will discuss how to fit the model using a Gibbs sampler, Variational Bayes, and geometric approaches. We then go over how to compute the metrics we introduced earlier in order to evaluate the results from these methods.

4.4.1 Model

The model that we are interested in fitting is the following. First, we draw $\pi, \theta_k, \beta_m$ for $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, M$ from Dirichlet distributions with $\alpha_\pi, \alpha_\theta$, and $\alpha_\beta$ as their respective hyperparameters. Next, we pick a path, $z_m$, according to the categorical distribution defined by $\pi$. Finally, we generate our observations, $x_m$, from a Multinomial distribution with probability:

$$\sum_j \beta_{m,j} \theta_{\varphi_j^m}.$$  

This is the convex combination of vertices. More specifically, the model is the following:

$$\pi \sim \text{Dir}(\cdot \mid \alpha_\pi),$$
$$\theta_k \overset{iid}{\sim} \text{Dir}(\cdot \mid \alpha_\theta), \quad k = 1, 2, \ldots, K,$$
$$\beta_m \overset{iid}{\sim} \text{Dir}(\cdot \mid \alpha_\beta), \quad m = 1, 2, \ldots, M$$

(4.6)

$$z_m = i \mid \pi \overset{iid}{\sim} \text{Cat}(\cdot \mid \lambda_1, \lambda_2, \ldots, \lambda_I),$$

$$x_m, \theta_1, \theta_2, \ldots, \theta_K, \beta_m, \varphi, z_m \sim \text{Multi}
\left(\cdot \mid N_m, \sum_j \beta_{m,j} \theta_{\varphi_j^m}\right), \quad m = 1, 2, \ldots, M.$$  

Alternatively, the traditional way to view this model is by using a variable to indicate the level, i.e. $\zeta_m \in \{1, 2, \ldots, J\}^N$. Then, before we generate our observation, we first draw $\zeta_{m,n}$ according to a categorical distribution with probability $\beta_m$. The observation is created from the categorical distribution with the vertex identified by the path, $\varphi_{\zeta_m}$, and the level, $\zeta_{m,n}$. This vertex is denoted as
The model can be alternatively expressed as following.

\[
\begin{align*}
\pi & \sim \text{Dir}(\cdot | \alpha_{\pi}), \\
\theta_k & \overset{iid}{\sim} \text{Dir}(\cdot | \alpha_{\theta}), \quad k = 1, 2, \ldots, K \\
\beta_m & \overset{iid}{\sim} \text{Dir}(\cdot | \alpha_{\beta}), \quad m = 1, 2, \ldots, M \\
zm & = i | \pi \overset{iid}{\sim} \text{Cat}(\cdot | \lambda_1, \lambda_2, \ldots, \lambda_K), \quad m = 1, 2, \ldots, M \\
\zeta_{m,n} & = k | \beta_m \overset{iid}{\sim} \text{Cat}(\cdot | \beta_m), \quad m = 1, 2, \ldots, M \quad n = 1, 2, \ldots, N_m
\end{align*}
\] (4.7)

4.4.2 Gibbs update

While the traditional approach is to sample from a collapsed version of (4.7), we wish to discuss how to sample from Model (4.6). We sample from that model so that our Gibbs sampler will be in the same spirit as the geometric approach. Then, the Gibbs updates are the following if we suppress the hyperparameters:

- For \( \theta_k \mid \{ \beta_m \}_{m=1}^{M}, \{ \theta_{k'} \}_{k'=1, k' \neq k}^{K}, \pi, \Phi(T) \) for \( k = 1, 2, \ldots, K \), propose a new \( \theta'_k \) with probability function \( g(\cdot | \cdot) \) and accept it with the following acceptance probability:

\[
\frac{\Pi_{m: k \in \varphi^m} \text{Multi}(x_m, N_m; \beta_{m,(k)} \theta'_k + \sum_{j \neq k} \beta_{m,j} \theta_{\varphi^m j}) \text{Dir}(\theta'_k | \alpha_{\theta})}{\Pi_{m: k \in \varphi^m} \text{Multi}(x_m, N_m; \sum_j \beta_{m,j} \theta_{\varphi^m j}) \text{Dir}(\theta_k | \alpha_{\theta})} \frac{g(\theta_k | \theta'_k)}{g(\theta'_k | \theta_k)}. \tag{4.8}
\]

- For \( \beta_m \mid \{ \theta_k \}_{k=1}^{K}, \varphi^m, \pi \) for \( m = 1, 2, \ldots, M \), propose a new \( \beta'_m \) with probability function \( g(\cdot | \cdot) \) and accept it with the following acceptance probability:

\[
\frac{\text{Multi}(x_m, N_m; \sum_j \beta_{m,j} \theta_{\varphi^m j}) \text{Dir}(\beta'_m | \alpha_{\beta})}{\text{Multi}(x_m, N_m; \sum_j \beta_{m,j} \theta_{\varphi^m j}) \text{Dir}(\beta_m | \alpha_{\beta})} \frac{g(\beta_m | \beta'_m)}{g(\beta'_m | \beta_m)}. \tag{4.9}
\]

- For \( z_m \mid \{ \theta_k \}_{k=1}^{K}, \beta_m, \pi, \Phi(T) \) for \( m = 1, 2, \ldots, M \), set \( z_m = i, i = 1, 2, \ldots, I_\rho \), with the
following probability:

$$\pi_i \text{Multi} \left( x_{m, i} \mid N_m, \sum_j \beta'_{m, j} \theta_{x_{m, j}} \right). \quad (4.10)$$

- For $\pi \mid z$, draw $\pi$ from $\text{Dir}(\bar{\alpha}_\pi)$ where for $i = 1, 2, \ldots, I$:

$$(\bar{\alpha}_\pi)_i = (\alpha_\pi)_i + \sum_m \mathbb{I}(z_m = i).$$

From this description, we see that our approach is better able to take advantage of the model and propose more global moves. Because $\beta_m$ and $z_m$ are independently drawn for each document and $\theta_k$ is independently drawn, we see that the proposed Gibbs sampler has two steps that only affect individual documents, one step that affects all documents, and one step that indirectly affects all documents. As a result, even the two most "local" steps affect the entire document affect the entire document. Not only that, we still are able to sample the entire support of these variables.

### 4.4.3 Estimating using Variational Bayes

For Variational Bayes (VB), we will run it on the model specified in Model (4.7) because the likelihood factorizes in the following manner:

$$P(\{x_{m, i}\}_{m=1}^M, \{\zeta_{m, n}\}_{m=1}^M, z, \{\theta_k\}_{k=1}^K, \{\beta_m\}_{m=1}^M, \pi \mid \alpha_\beta, \alpha_\theta, \alpha_\pi, \Phi(T)) =$$

$$P(\{x_{m, i}\}_{m=1}^M \mid \{\zeta_{m, n}\}_{m=1}^M, z, \{\theta_k\}_{k=1}^K, \Phi(T)) \prod_{m=1}^M P(\zeta_m \mid \beta_m) \quad (4.11)$$

$$P(\{\beta_m\}_{m=1}^M \mid \alpha_\beta P(\{\theta_k\}_{k=1}^K \mid \alpha_\theta) P(z \mid \pi) P(\pi \mid \alpha_c)$$

While we defer the derivation to the appendix, we list the variational distributions in this lemma.

**Lemma 4.4.1.** Let $\ell(\cdot)$ denotes the level of a vertex. Then, for Model (4.7), the variational distributions under the mean field approximation are the following:

$$q(\beta_m) = \text{Dir}(\bar{\alpha}_\beta_m),$$

$$q(\theta_k) = \text{Dir}(\bar{\alpha}_\theta_m),$$

$$q(\pi) = \text{Dir}(\bar{\alpha}_\pi),$$

$$q(\zeta_{m, n} = j) \propto \exp \left( \psi((\bar{\alpha}_\beta_m)_j) - \psi \left( \sum_{j'}(\bar{\alpha}_\beta_m)_{j'} \right) + \right)$$
\[ \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{1} (x_{m,n} = d) \mathbb{1} (\zeta_{m,n} = \ell(V_k)) \mathbb{E}_{z_m} (\mathbb{1} (k \in \varphi^{z_m})) \left( \psi((\tilde{\alpha}_{\theta_k})_d) - \psi\left( \sum_{d'} (\tilde{\alpha}_{\theta_k})_{d'} \right) \right), \]

\[ q(z_m = i) \propto \exp \left( \psi((\tilde{\alpha}_\pi)_i) - \psi\left( \sum_{i'} (\tilde{\alpha}_\pi)_{i'} \right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{1} (x_{m,n} = d) \mathbb{1} (k \in \varphi^i) \mathbb{E}_{\zeta_{m,n}} (\mathbb{1} (\zeta_{m,n} = \ell(V_k))) \left( \psi((\tilde{\alpha}_{\theta_k})_d) - \psi\left( \sum_{d'} (\tilde{\alpha}_{\theta_k})_{d'} \right) \right) \right). \]

Here,

\[ (\tilde{\alpha}_{\beta_m})_j = (\alpha_\beta)_j + \sum_{n=1}^{N} \mathbb{E}_{\zeta_{d,n}} (\mathbb{1} (\zeta_{m,n} = j)) \]

\[ j = 1, 2, \ldots, J \]

\[ m = 1, 2, \ldots, M \]

\[ (\tilde{\alpha}_{\theta_k})_d = (\alpha_\theta)_d + \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbb{1} (x_{m,n} = d) \mathbb{E}_{\zeta_{m,n}} (\mathbb{1} (\zeta_{m,n} = \ell(V_k))) \mathbb{E}_{z_m} (\mathbb{1} (k \in \varphi^{z_m})) \]

\[ d = 1, 2, \ldots, D \]

\[ (\tilde{\alpha}_\pi)_i = (\alpha_\pi)_i + \sum_{m=1}^{M} \mathbb{E} (\mathbb{1} (z_m = i)) \]

\[ i = 1, 2, \ldots, I_\rho. \]

We now make a few notes about this. First, we denote this algorithm as Tree Variational Bayes (Tree VB). Because we use the mean field approximation, conjugacy will allow us to derive closed form updates. From this derivation, there are a few computation shortcuts, particularly for higher dimensional data. Notice that the probability of a word’s level only depends on the word itself. This simplifies the computation of \( q(\zeta_{m,n}) \) because we only need to compute the probability of the levels for each word. Then, the probability of all words can be computed through matrix multiplication of the total probability of vertices across all paths arranged by level and a matrix where each row is the expectation of the log of \( \tilde{\alpha}_{\theta_k} \) for \( k = 1, 2, \ldots, K \). This same idea can be applied to calculate \( q(z_m) \) and \( \tilde{\alpha}_{\theta_k} \). Using the probability of the levels for each word and the total word count instead of the individual words with each document, we can perform matrix operations to compute the distribution of \( z_m \) and \( \theta_k \). Finally, the last piece we need is the lower bound, which we give in the following lemma.
Lemma 4.4.2. The evidence lower bound is

\[
\mathbb{E} \left( \log(P(x \mid z, \zeta, \theta)) \right) + \mathbb{E} \left( \log(P(\zeta \mid \beta)) \right) + \mathbb{E} \left( \log(P(\theta \mid \alpha_\theta)) \right) + \mathbb{E} \left( \log(P(\beta \mid \alpha_\beta)) \right) + \\
\mathbb{E} \left( \log(P(z \mid \pi)) \right) + \mathbb{E} \left( \log(P(\pi \mid \alpha_\pi)) \right) - \mathbb{E} \left( \log(q(z)) \right) - \mathbb{E} \left( \log(q(\zeta)) \right) - \mathbb{E} \left( \log(q(\theta)) \right) - \\
\mathbb{E} \left( \log(q(\beta)) \right) - \mathbb{E} \left( \log(q(\pi)) \right),
\]

where

\[
\mathbb{E} \left( \log(P(x \mid z, \zeta, \theta)) \right) = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I} (x_{m,n} = d) \mathbb{E} \left( \mathbb{I} (\ell(V_k) = \zeta_{m,n}) \right) \mathbb{E} \left( \mathbb{I} (k \in \varphi^m) \right) \\
\left( \psi((\tilde{\alpha}_{\theta_k})_d) - \psi \left( \sum_{d'} (\tilde{\alpha}_{\theta_k})_{d'} \right) \right),
\]

\[
\mathbb{E} \left( \log(P(\zeta \mid \beta)) \right) = \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{E} \left( \mathbb{I} (\zeta_{m,n} = j) \right) \left( \psi((\tilde{\alpha}_{\beta_m})_j) - \psi \left( \sum_{j'} (\tilde{\alpha}_{\beta_m})_{j'} \right) \right),
\]

\[
\mathbb{E} \left( \log(P(z \mid \pi)) \right) = \sum_{m=1}^{M} \sum_{i=1}^{I} \mathbb{E} \left( \mathbb{I} (c_d = i) \right) \left( \psi((\tilde{\alpha}_{\pi})_i) - \psi \left( \sum_{i'} (\tilde{\alpha}_{\pi})_{i'} \right) \right),
\]

\[
\mathbb{E} \left( \log(P(\theta \mid \alpha_\theta)) \right) = \sum_{k=1}^{K} \sum_{d=1}^{D} \left( (\alpha_\theta)_d - 1 \right) \left( \psi((\tilde{\alpha}_{\theta_k})_i) - \psi \left( \sum_{i'} (\tilde{\alpha}_{\theta_k})_{i'} \right) \right) - \\
K \log(B(\alpha_\theta)),
\]

\[
\mathbb{E} \left( \log(P(\beta \mid \alpha_\beta)) \right) = \sum_{m=1}^{M} \sum_{j=1}^{J} \left( (\alpha_\beta)_j - 1 \right) \left( \psi((\tilde{\alpha}_{\beta_m})_j) - \psi \left( \sum_{j'} (\tilde{\alpha}_{\beta_m})_{j'} \right) \right) - \\
M \log(B(\alpha_\beta)),
\]

\[
\mathbb{E} \left( \log(P(\pi \mid \alpha_\pi)) \right) = \sum_{i=1}^{I} \left( (\alpha_\pi)_i - 1 \right) \left( \psi((\tilde{\alpha}_{\pi})_i) - \psi \left( \sum_{i'} (\tilde{\alpha}_{\pi})_{i'} \right) \right) - \log(B(\alpha_\pi)).
\]

The other expectations can be found using Lemma 4.4.1.

Again, there are a few computational shortcuts. If we use the same prior as the model, we do not need to compute \( \mathbb{E} \left( \log(P(\theta \mid \alpha_\theta)) \right), \mathbb{E} \left( \log(P(\beta \mid \alpha_\beta)) \right), \mathbb{E} \left( \log(P(\beta \mid \alpha_\beta)) \right), \) and their VB counterparts because they cancel each other out. Then, for \( \mathbb{E} \left( \log(P(x \mid z, \zeta, \theta)) \right) \) and \( \mathbb{E} \left( \log(P(\zeta \mid \beta)) \right) \), the parts not in the digamma function are already computed when updating \( \tilde{\alpha}_{\theta_k} \) and \( \tilde{\alpha}_{\beta_m} \) respectively from Lemma 4.4.1 for \( k = 1, 2, \ldots, K \) and \( m = 1, 2, \ldots, M \). Hence, once the respective prior are subtracted from these quantities, it is possible to build an matrix, element wise multiply it, and then sum over it to compute these quantities.
4.4.4 Geometric algorithms

In the previous subsections, we have discussed standard sampling approaches. An interesting follow-up question is whether it is possible to use a geometric approach with theoretical guarantees to sample from this model. To begin this discussion, let us make the following additional assumption:

(A10) for any \( k, k' \in 1, 2, \ldots, K \), \( \|\theta_k - \theta_{k'}\| > \epsilon \) for some \( \epsilon > 0 \).

Because of these assumptions, the simplices defined by a path do not overlap and the topics are sufficiently distinct.

For our objective function, we cannot use the tree-based and path-based metrics because these metrics compare trees. Instead, we will use what we call the projected simplicial distance, which we define below.

Definition 4.4.1. Suppose that \( \theta_1, \theta_2, \ldots, \theta_K \) are points in \( \Delta^{D-1} \) and \( T \) denotes a tree. Given \( \eta \in \Delta^{D-1} \), let \( \lambda(\eta, G_i) \) denote the projection of \( \eta \) onto the simplex defined by \( \phi^i \) for \( i = 1, 2, \ldots, I_p \). Then, define the projected simplicial distance as following:

\[
d_\lambda(\eta, T, \theta_1, \theta_2, \ldots, \theta_K) = \min_i \|\eta - \lambda(\eta, G_i)\|.
\]

If it is clear which vertices are referred, we will denote the projected simplicial distance as \( d_\lambda(\eta, T) \).

The goal is to minimize the total simplicial distance. Note that the projected simplicial distance is similar to the path-based metric calculated between our estimated vertices and the data. However, the comparison for projected simplicial distance is one-sided in that we find the distance from the data point to the closest polytope. Further, instead of taking the max, we sum over these distances for the projected simplicial distance. In the noiseless case, this relationship can be made explicit through the following lemmas.

Lemma 4.4.3. Suppose that \( \theta_1, \theta_2, \ldots, \theta_K \) are points in \( \Delta^{D-1} \) and \( T \) denotes a tree. Let \( S_{[n]}^1, S_{[n]}^2, \ldots \) be a collection of noiseless observations generated from these topics and trees and \( \tilde{\eta}_m = \frac{1}{N_m} S_{[n]}^m \) for \( m = 1, 2, \ldots \). Let \( \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_K \) and \( \tilde{T} \) denote the estimated topics and tree respectively. If \( d_\rho(T, \tilde{T}) = 0 \), then \( d_\lambda(\tilde{\eta}_m, T, \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_K) = 0 \) for \( m = 1, 2, \ldots \).

Proof. Let \( \tilde{G}_i \) denote the estimated polytope defined by path \( \phi^i \) for \( i = 1, 2, \ldots, I_p \). Because \( d_\rho(T, \tilde{T}) = 0 \), then \( G_i = \tilde{G}_{\sigma_\rho(i)} \) for some path-preserving permutation, \( \sigma_\rho \), and all \( i = 1, 2, \ldots, I_p \). Due to the observations being noiseless, \( \tilde{\eta}_m \) must be in some \( G_i \) for \( m = 1, 2, \ldots \) and \( i = 1, 2, \ldots, I_p \). Consequently, \( d_\lambda(\tilde{\eta}_m, T) = 0 \) for \( m = 1, 2, \ldots \). \( \square \)
However, the reverse does not hold. Instead, we need to include a condition on the volume of each polytope. Let \( \text{Vol}(G) \) denote the volume of polytope \( G \). Then, we have the following lemma.

**Lemma 4.4.4.** Suppose that \( \theta_1, \theta_2, \ldots, \theta_K \) are points in \( \Delta^{D-1} \) and \( T \) denotes a tree such that each node except the leaf nodes has at least two children. Assume that Assumption (A1) and (A4) – (A9) holds. Let \( S^1_{[n]}, S^2_{[n]}, \ldots \) be a collection of noiseless observations generated from these topics and trees and \( \hat{\eta}_m = \frac{1}{N_m} S^m_{[n]} \) for \( m = 1, 2, \ldots \). Define \( \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_K \) and \( \tilde{T} \) denote the estimated topics and tree respectively. Further, let \( \hat{G}_i \) denote the estimated polytope defined by path \( \varphi^i \). If

\[
\bullet \quad d_{\lambda}(x_m, T) = 0 \quad \text{for} \quad m = 1, 2, \ldots \quad \text{and}
\]

\[
\bullet \quad \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{T} = \arg\min_{T', \theta_1', \theta_2', \ldots, \theta_K'} \sum_i \text{vol}(\hat{G}_i),
\]

then \( d_P(T, \hat{T}) = 0 \).

**Proof.** We prove this by construction. First, because we are in the noiseless case and assumptions (A1) and (A4) to (A8) hold, this tree is identifiable.

Consider the points generated from \( G_i \). Because of identifiability and assumptions (A9), these points can only be generated from \( G_i \). As there is infinite data, Assumption (A5) means that there exists no set \( A \) such that \( A \subset G_i \) and \( \hat{\eta}_m \notin A \) for \( m = 1, 2, \ldots \). Consequently, if \( d_{\lambda}(x_m, T) = 0 \) for all \( x_m \) generated from \( G_i \), the smallest such polytope must be \( G_i \). This must hold for any \( G_i, i = 1, 2, \ldots, I_\rho \). As a result, if the two conditions hold, \( d_P(T, \hat{T}) = 0 \).

Based on this lemma and the constructive proof, an algorithm might aim to reduce the simplicial distance while controlling for the volume. As a result, a proposed algorithm is the following:

1. Given \( \{S^m_{[n]}\}_{m=1}^M \), compute \( \hat{\eta}_m = \frac{1}{N_m} S^m_{[n]} \) for \( m = 1, 2, \ldots, M \).
2. Run spectral clustering with \( I_\rho \) clusters to assign \( \{S^m_{[n]}\}_{m=1}^M \). As before, let \( z \) indicate the assignment.
3. For each path, \( i = 1, 2, \ldots, I_\rho \), apply a geometric algorithm to learn \( J \) vertices, \( \hat{\theta}_{i,1}, \hat{\theta}_{i,2}, \ldots, \hat{\theta}_{i,J} \) from all \( \{\hat{\eta}_m\}_{m=1}^M \) such that \( z_m = i \). Let \( C \) denote the collection of topics learned by setting \( C_i = \{\hat{\theta}_{i,j}\}_{j=1}^J \) for \( i = 1, 2, \ldots, I_\rho \). Geometric algorithms include GDM [76], VLAD [75], or a hybrid version discussed in the appendix C.2.2.3.
4. Apply the tree merge procedure on \( C_i \) to get \( \tilde{\theta}_k \) for \( k = 1, 2, \ldots, K \).
5. Set \( z_m = \arg\min_{i \in 1, 2, \ldots, I_\rho} \|\hat{\eta}_m - \lambda(\hat{\eta}_m, G_i)\| \) for \( m = 1, 2, \ldots, M \).
6. Repeat steps 2 to 4 until the projected simplicial distance is small.
We denote this algorithm as a GeoTree with the geometric method used to learn the extreme points included next in parentheses. Clearly, the first two steps reduce the simplicial distance. We will show how the last step has a stability property while merging the estimated vertices into vertices of a tree.

To discuss our merge procedure, we will first describe it conceptually before stating the algorithm. We begin with a few definitions.

**Definition 4.4.2.** Given a set of paths, \( \mathcal{I} \), their collection of available extreme points indices, \( \mathcal{J}_i \) for \( i \in \mathcal{I} \), and a size, \( O \in \mathbb{N} \), define \( M_o = \{ i \in \mathcal{I}, j \in \mathcal{J}_i \} \) for \( o = 1, 2, \ldots, O \). Call \( M \) a merge group if \( (M_s)_1 \neq (M_o')_1 \) for any \( o, o' \in 1, 2, \ldots, O \).

At a high level, we want to merge one item from each group. The merge group indicates which items should be merged. Then, for a tree with \( K \) vertices, we need a set of \( K \) merge groups. This leads to the following definition.

**Definition 4.4.3.** Let \( \Xi \) be a set of \( K \) merge groups. Let \( (\Xi_k)_1 = \{ ((\Xi_k)_l)_1 \}_{l=1}^{\Xi_k} \) Suppose that for \( k, k', k'' \in 1, 2, \ldots, K \),

- \( |\Xi_k| = I_k \) where \( I_k \) is the number of paths that go through \( V_k \),
- \( (\Xi_k)_1 \subset (\Xi_{k'})_1 \) if \( V_{k'} \) the parent of \( V_k \), and
- \( (\Xi_k) \cap (\Xi_{k''}) = \emptyset \).

Then, \( \Xi \) is a tree set of merge groups.

The goal of the tree merge procedure is to minimize the following:

\[
\sum_k \max_{(i,j), (i',j') \in \Xi_k} \left\| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \right\|.
\]

(4.12)

The following lemma shows the benefit of doing so.

**Lemma 4.4.5.** Suppose that Assumption (A10) holds. For \( i = 1, 2, \ldots, I_\rho \), let \( \{ \hat{\theta}_{i,j} \}_{j=1}^{J} \) be the sets of estimates of the vertices of the polytope defined by \( \varphi^i \) from some procedure such that each vertex is estimated and if \( \hat{\theta}_{i,j} \) is an estimate of \( \theta_k \), \( \left\| \hat{\theta}_{i,j} - \theta_k \right\| < \frac{\epsilon}{4} \). Set \( \Xi_T \) to be the \( \Xi \) that minimizes (4.12). If \( \bar{\theta}_k = \sum_{(i,j) \in \Xi_{\tau_k}} \frac{1}{|\Xi_{\tau_k}|} \hat{\theta}_{i,j} \) for \( k = 1, 2, \ldots, K \), then \( \left\| \bar{\theta}_k - \theta_k \right\| < \frac{\epsilon}{4} \).

**Proof.** Under our assumptions about the vertices and estimated vertices, \( \left\| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \right\| < \frac{\epsilon}{2} \) if \( \hat{\theta}_{i,j}, \hat{\theta}_{i',j'} \) are estimates of \( \theta_k \) for some \( k = 1, 2, \ldots, K \) and \( \left\| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \right\| > \frac{\epsilon}{2} \) otherwise for all \( i \in 1, 2, \ldots, I_\rho \) and \( j = 1, 2, \ldots, J \). Further, if there are \( I_k \) paths that go through \( V_k \), there
are $l_k$ estimates of the vertex at that node because for each polytope defined by a path that goes through $V_k$, the procedure must return one estimate of each vertex from that polytope. Finally, this also demonstrates that if $\max_{(i,j),(i',j') \in \Xi_k} \| \tilde{\theta}_{i,j} - \tilde{\theta}_{i',j'} \|$ is minimized, estimates of vertices from polytopes defined by paths that share nodes will be coupled. Due to the shared nodes being at the top of the tree, the paths of a merge group for $V_k$ minimizing that quantity must be the subset of the paths of the merge group for $V_k$’s parents. As a result, minimizing (4.12) leads to estimates of the same vertices being grouped together and the set of merge groups being a tree set of merge group.

By triangle inequality, we have that

$$\| \tilde{\theta}_k - \theta_k \| = \left\| \sum_{i : k \in \varphi^i} \frac{1}{I_k} \left( \tilde{\theta}_{i,j} - \theta_k \right) \right\| = \frac{1}{I_k} \left\| \sum_{i : k \in \varphi^i} \tilde{\theta}_{i,j} - \theta_k \right\| < \frac{1}{I_k} \sum_{i : k \in \varphi^i} \left\| \tilde{\theta}_{i,j} - \theta_k \right\| < \frac{\epsilon}{4}.$$  

We now make a few comments about this procedure. First, this merge procedure does not affect the convergence rate of estimating $\theta_k$. Indeed, this algorithm can help the convergence rate for all nodes except the leaf nodes. As a result, if identifiability holds and we make the appropriate assumptions, the convergence rate for affine tree geometric algorithm depends on the convergence rates for identifying the affine space and estimating the vertices in that space. Next, while merging vertices might increase the projected simplicial distance, the maximum projected simplicial distance for a data point remains bounded. Finally, it remains challenging to discuss precisely how it affects the volume and simplicial distances. The projected simplicial distance depends on understanding the overlap between the changed estimated polytope and the data and the projected distance for the remaining points. Changing one point can affect this by potentially pulling the estimated polytope out of the same affine space. This is challenging to characterize. However, notice that the proof in Lemma 4.4.5 holds for any convex combination of the points in the tree set. It is possible to test different combinations and find the ones that minimize either the volume or simplicial distance. However, empirical evidence seems to indicate that this is not necessary.

However, finding $\Xi$ that minimizes (4.12) is challenging. Instead, we use the following greedy algorithm that we call the tree merge procedure.

- Let $J_i = \{1, 2, \ldots, J\}$ for $i = 1, 2, \ldots, I_\rho$. Find $M^1$ of size $I_\rho$ for $\{1, 2, \ldots, I_\rho\}$ and all $J_i$ using single linkage such that $\max_{(i,j),(i',j') \in M^1} \| \tilde{\theta}_{i,j} - \tilde{\theta}_{i',j'} \|$ is minimized. For each $i, j \in M^1$, remove $j$ from $J_i$.

- Let $\tilde{K}_2$ be the set of nodes at the second level and let $I_k$ denote the number of paths that go through node $k \in \tilde{K}_2$. 

89
For the $k$ with the largest $I_k$, find $\mathcal{M}^k$ of size $I_k$ for all paths and all $\mathcal{J}_i$ using single linkage such that $\max_{(i,j),(i',j') \in \mathcal{M}^k} \| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \|$ is minimized. Again, remove $j$ from $\mathcal{J}_i$ for each $i, j \in \mathcal{M}^k$, $k = 1 \in \mathcal{K}_2$. Record the paths as associated with $\mathcal{M}^k$ as $P_k$.

For the $k'$ with the next largest $I_k'$, find $\mathcal{M}^{k'}$ of size $I_k'$ for $\{1, 2, \ldots, \mathcal{I}_\rho\} \setminus P_k$ and their associated $\mathcal{J}_i$'s using single linkage such that $\max_{(i,j),(i',j') \in \mathcal{M}^{k'}} \| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \|$ is minimized. Again, remove $j$ from $\mathcal{J}_i$ for each $i, j \in \mathcal{M}^{k'}$, $k = 1 \in \mathcal{K}_2$. Record the paths as associated with $\mathcal{M}^{k'}$ as $P_{k'}$.

Repeat this for the next largest $I_k''$ except for $\{1, 2, \ldots, \mathcal{I}_\rho\} \setminus P_k \cup P_{k'}$ and their associated $\mathcal{J}_i$'s and so on until all nodes at the second level have groups assigned to them.

• For each node, $V_k$, in the previous step, repeat the previous step for its children. Instead of starting with all paths, begin with the paths in $P_k$.

• Repeat this for the children’s children and so on until merge groups at level $J - 1$ have been identified. The unassigned vertices are associated with the nodes at the leaf.

• Return $\bar{\theta}_k = \sum_{\mathcal{M}^k} \frac{1}{|\mathcal{M}_k|} \hat{\theta}_{i,j}$ for $k = 1, 2, \ldots, K$.

The algorithm is greedy because the use of single linkage at each step minimizes the following:

$$\max_{\{i,j\},\{i',j'\} \in \mathcal{M}} \| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \|_2.$$ 

Further, we minimize the quantity from the root vertex to the vertices at the next level, the vertices below that, and so on. The advantage of such an approach is that it maintains the tree structure as nodes are merged. We can also establish the following corollary via induction on the nodes, starting from the root node. Because of length, we move the proof to the appendix.

**Corollary 4.4.1.** Assume the assumptions of Lemma 4.4.5 hold. Then, the the algorithm discussed above returns a tree set of merge groups, $\Xi_T$. Further, if $\hat{\theta}_k = \sum_{(i,j) \in \Xi_T,k} \frac{1}{|\Xi_T,k|} \hat{\theta}_{i,j}$ for $k = 1, 2, \ldots, K$, then $\| \hat{\theta}_k - \theta_k \| < \frac{\epsilon}{4}$.

**4.4.5 Metric computation**

It can be difficult to compute the path and tree metric because calculating these distances requires finding the right combination of nodes or polytopes that minimizes the $L_2$ distance or the Hausdorff metric respectively. While it reduces the number that need to be considered, the definition of the path and tree permutation do complicate this minimization. As a result, we discuss how to compute
the path and tree metric in a greedy manner and for a balanced tree. In other words, all nodes except the children node have a certain number of children, $K$.

First, because it is easier, we discuss how to compute the tree-based metric. Suppose that we are given $\theta_1, \theta_2, \ldots, \theta_K$ associated with tree $T$ and $\theta'_1, \theta'_2, \ldots, \theta'_K$ associated with tree $T'$. Further, suppose that for $T$ and $T'$, $J = J'$ and $K = K'$. Again, let $\ell(\cdot)$ denote the level function. Denote the estimated tree base metric as $\hat{d}_T$. Then, the metric can be calculated in the following recursive manner.

1. Set $\hat{d}_T = 0$ and $j = 1$.
2. Take the vertex associated with the root node of the tree. Increase $\hat{d}_T$ by $\frac{\mathcal{I}_\rho}{K^{j-1}} \| \theta_1 - \theta'_1 \|$.
3. Find a permutation $\sigma$ such that $\sigma = \min_\sigma \sum_{V_k = j+1} \| \theta_k - \theta'_{\sigma(k)} \|$.
4. Increase $j$ by 1. For each $k$ such that $\ell(V_k) = j$, treat $V_k$ and $V_{\sigma_j(k)}$ as the "root" nodes and consider their "subtrees". Repeat the previous steps for these "subtrees".
5. Repeat the previous step with every children and its permuted counterpart node as the "root" nodes and their respective "subtrees" and so on. Stop after $j = J - 1$. Return $\hat{d}_T$.

Through this procedure, we are computing at each level the following quantity:

$$\min_{\sigma_j} \sum_{k: \ell(V_k) = j} \frac{\mathcal{I}_\rho}{K^{j-1}} \| \theta_k - \theta'_{\sigma_j(\cdots(\sigma_2(k))))} \|.$$ 

In other words, we are minimizing at each level instead of minimizing across all permutations at the end.

Next, we discuss how to compute the path-based metric in a recursive manner in the set-up as above. Let $\hat{\sigma}_P$ denote the estimated path-preserving permutation.

1. Start at the level of the root node. Find $i$ and $i'$ such that $d_H(G_i, G'_{i'})$ is minimized. Set $\hat{\sigma}_P(i) = i'$.
2. Go down a level.
   - For each $k, k'$ such that $k \notin \varphi^i$, $k' \notin \varphi'^{i'}$ and $\ell(V_k), \ell(V'_{k'})$ equal to the current level, repeat the step above, but only among the paths that go from the root node and through that node.
   - For $k, k'$ such that $k \in \varphi^i, k' \in \varphi'^{i'}$, repeat the step above for each of the nodes’ children except any node on a path for which $\hat{\sigma}_P$ is defined. Again, consider only the paths that go from the root node through a particular child node.
3. Repeat the step above until all paths have been matched. Return \( \sum_{i \in I_p} d_H(G_i, G_{\sigma_p(v)}) \).

This procedure is a greedy search through the node’s children for the next valid match with the smallest Hausdorff distance between the simplices defined by the paths in order to estimate the path-preserving permutation. Again, this does not compute the exact path-based metric. It is possible that the sum of the Hausdorff distance between the simplices defined by a subset of paths and another subset might be smallest, but one subset might get matched to some other subset because there might be a path from each subset such that the Hausdorff distance between the simplices is smaller than the Hausdorff distance computed from the optimally matched subsets.

### 4.5 Simulation and data analysis

#### 4.5.1 Simulation

We ran a simulation study to understand how our methods performed. In particular, we were interested in our methods’ performance as the dimension of the simplex increased and the size of the tree increased. The performance was evaluated according to the path and tree metric that we defined. In order to also compare our methods against existing methods for Latent Dirichlet Analysis, we also computed a minimized sum of \( L_2 \) distances between vertices. Another purpose in running these simulations was to compare how fast our code would run.

To that end, we had two sets of experiments. First, we examined the effect of the dimension of the simplex with \( D \) set to 20, 50, 100, and 2000. The tree we used had a depth three and all nodes except the leaf nodes having two children nodes. We generated these vertices from a Dirichlet distribution with \( \alpha = 1 \). Next, to ensure that all paths would be used, we set the path probability to be 0.35, 0.25, 0.25, and 0.15. Finally, we generated the level probability from another Dirichlet distribution with \( \alpha = 1 \). Using this, we generated 1000 observations of size 1000 according to the model specified in Section (4.4.1). The other set of experiments was examining the effect of increasing the size of the tree. To that end, we separately tested increasing the depth of the tree to five and the number of children to four. We again set the path probability with two equal to 0.1, four equal to 0.075, and the remaining equal to 0.05. To generate the level probability and the vertices, we used an appropriate Dirichlet with \( \alpha = 1 \). For this experiment, we generated 2000 observations of size 2000 according to the model specified in Section (4.4.1) so that there would be at least 100 observations per path on expectation.

Then, we fitted these experimental data sets using Gibbs LDA, the Gibbs method outlined in Section 4.4.2, Tree VB outlined in Section 4.4.3, and the geometric methods discussed in Section 4.4.4. For the Gibbs algorithms, we ran 2000 iterations and used the SALTSampler to propose a new \( \theta_k \) and \( \beta_m \) [19]. While it is possible to check a new \( \beta_m \) after proposing a change to one of
Table 4.1: Various distances between vertices used to generate the simulation and vertices as estimated by different methods for data of distinct dimension and in parentheses, the average projected simplicial distance with the minimum distance across all methods in bold. The first line presents the minimized sum of $L_2$ distance between topics. The next two lines show the path and tree metric respectively defined in (4.3.2) and (4.3.4) and computed in the manner discussed in Section 4.4.5. The last line illustrates the projected simplicial distance (4.4.1). The table on the left shows results when the tree size is fixed so that the tree is three deep and each node except the leaf nodes has two children and $D$ varies whereas the table on the right displays the results when $D$ is fixed to be 100 and the tree size varies.

its dimension, this is too consuming for $\theta_k$. Instead, we checked the acceptance probability for $\theta_k$ after proposing changes to $\frac{D}{10}$ dimensions. Meanwhile, we set the seed to be the same so that all geometric methods had the same initialization. Finally, for VB, we ran it for 100 iterations or until the ELBO stopped improving by 1e-3 depending on whichever happened first.

We now discuss the results of the experiments from Table 4.1, C.1, and C.2. First, as $D$ increases, the overall performance of our methods appears to improve according to the metrics. However, this is because of the constraint that the proportions need to sum up to one and the curse of dimensionality. As these metrics are based on the $L_2$ distance, the curse of dimensionality means that points aren’t as far or close to each other in higher dimensions. Further, due to the proportions needing to sum up to one, the proportions live in $L_1$ space. This becomes increasingly smaller in
L₂ space as the dimension grows. Despite this caveat, the geometric methods perform best and are significantly faster. On the other hand, as the tree grows in the number of children or depth, performance suffers across all methods. This makes sense because there are more simplices or simplices of higher dimension to fit. With the exception of a tree with depth three and each node having four children, the geometric algorithm performs well according to our metric at a fraction of the time compared to other methods. However, in that scenario, one geometric method is still better able to estimate the vertices if we ignore the structure and another geometric method is close in performance according to the tree-based metric. Finally, it appears that we can model select according to the projected simplicial distance defined in (4.4.1). While we cut this distance to the third decimal point, expanding it to the fourth shows that in most cases, the method with the smallest projected simplicial distance also performs best according to the other metrics. This is useful because while we may not be able to compare a method’s estimated vertices against the vertices used to generate the data, we can always compare the estimated vertices against the data itself.

While it appears the hybrid version of the algorithm works best in our experiments, other geometric algorithms may have similar performance. This motivated us to understand how the different geometric algorithms work. Starting from the same, but different initialization, the plots in Figure 4.2 show a case in which the hybrid method does not work as well as method using VLAD or GDM. Here, D = 20 and the tree has depth three with each node except the leaf nodes having two children nodes. From these plots, we see that VLAD gradually adjusts the estimated vertices. As more data are correctly assigned to a particular path, the estimated simplices rotate and expand to
more closely resembles the simplex used to generate the data. Meanwhile, the vertices estimated by GDM can undergo more drastic changes. Because more data are near the shared vertex, this can help GDM identify these vertices and consequently, the simplex used to generate the data. The hybrid method does not perform as well because it appears stuck in a local minimum. In other words, for two of the simplices identified by the algorithm, extending from the overall center through the centers found by VLAD and to the edge of the data result in similar vertices being identified. As a result, the initialization avoids this by using the data to identify the centers to stretch through. Consequently, it is vital to try various starts. Indeed, while we do not report these runs, the geometric algorithms can perform better according to the path-based metric in the scenario of a tree with depth three and each node having four children.

4.5.2 New York Times corpus

We decided to use the New York Times Corpus for a real world exploration. The annotated set contains over 1.8 million documents from 1987 to 2007. However, due to memory issues, we examined a subset of articles from 2000. First, we removed all English stopwords from the SMART information retrieval system and from the stopwords function with "en" as its argument of the Rtm package. We then reduced our vocabulary size by examining the 1000 most common words across all articles from 2000. Then, we kept 30% of the articles that had more than 100 of these words from each month. This left us with 12,779 articles to analyze.

We again sought to apply the algorithms introduced in this paper using the same initializations and settings discussed in the previous section. However, because of the time the algorithms took, we could only compare the geometric methods and Variational Bayes. For a tree of depth four and each node except the leaf node having three children, the Gibbs sampler needed around eight hours to fit the best tree-based mixture and GeoTree (GDM) took slightly longer than eight hours to fit five tree-based mixtures: 3 2, 3 4, 4 3, 6 2, and 3 6. Notice that under our assumptions, these mixtures are ordered from fewest to most simplices in the mixture. We then fit the best tree-based mixture and GeoTree (GDM) took slightly longer than eight hours to fit five tree-based mixtures: 3 2, 3 4, 4 3, 6 2, and 3 6. Notice that under our assumptions, these mixtures are ordered from fewest to most simplices in the mixture.

Besides presenting computational issues, this data set presented an additional challenge because the structure of the tree is unknown. All of the theory discussed in this paper and the simulations depended on the tree depth being known and the number of children per node being set ahead of time. To deal with this issue, we fit tree-based mixtures of the following sizes with the depth first and the number of children node second: 3 2, 3 4, 4 3, 6 2, and 3 6. Notice that under our assumptions, these mixtures are ordered from fewest to most simplices in the mixture.

We again sought to apply the algorithms introduced in this paper using the same initializations and settings discussed in the previous section. However, because of the time the algorithms took, we could only compare the geometric methods and Variational Bayes. For a tree of depth four and each node except the leaf node having three children, the Gibbs sampler needed around eight hours to fit the best tree-based mixture and GeoTree (GDM) took slightly longer than eight hours to fit five tree-based mixtures: 3 2, 3 4, 4 3, 6 2, and 3 6. Notice that under our assumptions, these mixtures are ordered from fewest to most simplices in the mixture.
model selected using AIC, BIC, and based on our simulations, the projected simplicial distance. Here, the log likelihood is computed with the last line of the model specified in (4.6).

These criteria led to the following findings. AIC and the projected simplicial distance favored a tree of depth six and each node except the leaf node having two children for both the GeoTree and Tree VB algorithms. Interestingly enough, while the geometric algorithm was the same, AIC and the projected simplicial distance did not recommend the same GeoTree run. This suggests that as seen in Figure 4.3, even though the AIC probability is computed in part by projecting the mean proportion to its assigned simplex, the $L_2$ ball may not correspond to the areas of equal probability under the multinomial distribution. Meanwhile, BIC preferred a tree of depth three and each node except the leaf node having five children for the GeoTree algorithm and for the Tree VB. This seems sensible that BIC prefers a less "complicated" model because it places a higher penalty on the number of parameters. Indeed, the number of parameters increases by the number of documents when the level of the tree goes up by one and by the number of words for each added topic. So while a deeper tree with more probability simplices is better able to explain the data due to more topics and more simplices for a document to belong to, the gain is not enough for BIC. It appears then that based on the structure of the preferred tree-based mixture, AIC and the projected simplicial distance prefers probability simplices that are larger in quantity and size whereas BIC appears to prefer more, but not too many, simpler probability simplices when the tree size is unknown. Next, the best Tree VB for each criterion outperforms the best GeoTree algorithm for the same criterion. As discussed in the appendix when comparing two fits, this might be because VB might be better able to handle model misspecification due to VB using a soft cut-off when updating the estimates.
for each topics’ alpha prior. The geometric methods use a hard cut-off since it must merge a certain number of estimated topics to estimate the topic associated with a node. Consequently, we defer a discussion of the geometric results to the appendix.

We now make a few comments on these results. First, across all of the best fits, the topic associated with the root node is similar. The top three words are "people", "years", and "time". "Year" is also within the top five words. This makes sense because after removing the stop words, news articles are usually about what people do at a given time point. Further, if we examine the top five words, there are sensible topics. For instance, Topic 10 in Figure 4.4 in the tree-based mixture of depth six and each node except the leaf node having two children appears to be about the performance of jazz music. However, while the topics associated with nodes lower in the tree are more specific, there are few common topics at each level below the first even if the same algorithm is used to estimate the tree. This makes sense because the tree-based mixtures preferred by AIC and BIC are nearly mirror opposites of each other. AIC prefers a "deep" tree with depth six, but few children, whereas BIC prefers a "shallow" tree with depth three, but many children.

Instead, one way to compare these trees is to find potentially common topics at the bottom of the tree-based mixtures. Then, we can look at the path associated with that topic and examine the related topics on the nearby paths. For instance, take Topic 58 from the tree-based mixture of depth six and each node except the leaf nodes having two children. It appears to about the candidates for the 2000 Presidential Election. Based on the top five words, Topic 21 from the tree-based mixture of depth three and each node except the leaf nodes having five children is most similar. The only word that differs is "Clinton" and "McCain". However, notice that in the the tree-based mixture of depth six and each node except the leaf nodes having two children, Topic 14 includes "Clinton". Because Topic 14 is on the same path and is a "grandparent" of Topic 58, the tree-based mixture of depth six and each node except the leaf nodes having two children shows that Topic 58 is a particular type of campaign. Indeed, the path with topics 3, 7, 14, 29, and 58 appear to be about the 2000 election campaigns. Due to there not being as many topics in the tree-based mixture of depth three and each node except the leaf node having five children, it is possible that Topic 21 is a summary of candidates in the 2000 elections. As further evidence of this, Topic 20 in the tree-based mixture of depth three and each node except the leaf node is about police officers and their investigation. This is similar to Topic 56, which also has Topic 14 as a "grandparent" in the tree-based mixture of depth six and each node except the leaf node.

However, there are differences in what other topics are tied to the tree-based mixtures of depth three and each node except the leaf node having five children. Topic 17 and 18 appear to be related to companies whereas Topic 19 appears to be about schools. However, they are related through the parent topic, Topic 4, which is about government officials, but includes words such as "state" and "federal". For instance, one article related to the path of 1, 4, and 19 is about state and federal
Figure 4.4: Best fitted topic tree showing the five most likely words associated with each topic from Variational Bayes according to the criterion listed below. The number of documents associated with each path is also displayed at the bottom of the tree.
education spending. Meanwhile, the topics from the tree-based mixture of depth six and each node except the leaf node closest to Topic 17, 18, and 19 are from a different part of the tree as the only common topic is the topic at the root node. Interestingly enough, these topics that are closest to Topic 17, 18, and 19 share a common great-grandparent "topic".

Finally, it appears that the tree-based mixtures preferred by AIC and BIC suffer from having repeated topics. For instance, the topic on Bush’s tax plan is Topic 44 in the tree-based mixture of depth six and each node except the leaf node having two children. Because Topic 44 is part of a path that does not include the nodes associated with Topics 3, 7, 14, 29, and 58, which is the path about campaigns, other top words include "Gore" and "president" and not Social Security. In other words, the topic needs to re-establish its relationship to election news. Indeed, as Topics 22 and 48 in the tree-based mixture of depth six and each node except the leaf node having two children demonstrate, there might be very similar topics at various levels that are needed to define these relationships. In this case, the tree-based mixture does not have enough children to relate these topics. On the other hand, there are multiple topics related to books and the presidential election in the tree-based mixture of depth three and each node except the leaf node having five children. For instance, take paths with Topics 1, 4, and 21 and Topics 1, 5, and 25. "Bush", "president", and "campaign" are all top words for topics along that path. As a result, both paths are concerned with the election campaigns, but one is associated with articles on former Secretary of State Clinton’s run for the New York Senate seat in 2000 whereas the latter is associated with the Republican presidential primary. It appears that the tree-based mixture in this scenario does not have enough depth to relate these paths because both these paths have similar topics and so may have more common ancestor topics if there were more topics.

4.6 Conclusion

In this chapter, we explored tree-based mixtures of probability simplices. These mixtures serve as a parsimonious extension of admixture models to account for further heterogeneity in the model because while allowing for multiple admixture models, the vertices are shared in a way that can be organized into a tree. Here, the vertex at the root node is shared by all simplices, the vertices at the next level are shared by subsets, and so on until the nodes at the leaf are unique to each simplex. While it was straightforward to demonstrate posterior consistency in distribution, we established when the structures are identifiable in the noiseless case and used these insights to discuss when the tree-based mixture structure is identifiable in the typical case. Of course, establishing identifiability and consistency in structure require metrics for the tree-based mixture. To that end, we introduced a path-based metric and a tree-based metric. Both involved defining a permutation and minimizing certain metrics against these permutation. Then, based on our theoretical understanding, we
proposed Gibbs, VB, and geometric approaches to sample from this model. We then applied these methods to simulation and found that in most cases, the geometric approaches provided the best fits according to the metrics we introduced earlier. In addition, these approaches were significantly faster than the other approaches. As a result, we used these methods and Tree VB to find the "best" tree-based mixture for a subset of New York Times articles from 2000. When we examined the topic tree, the organization was interpretable and understandable. Indeed, examining topics on the same path or in a subtree seemed to reflect particular sections of the newspaper or real world events. It is encouraging that the geometric and VB methods were able to discover such semantic relationships.
CHAPTER 5

Conclusion

This thesis lays the framework for modeling how data that lie in a simplex change over time given only the data itself. Chapter 2 discussed how to do so with an affine transformation and various "error" models. The models are summarized in this manner because they explain the difference between the data from one time point transformed and the data observed at the next time point. However, such models mostly assumed that the direction each data point moved in was uniformly distributed. A quick data exploration of the Los Angeles County income proportions from 1990 to 2010 revealed that the direction might be correlated based on the current proportions. The next chapter used ideas from directional statistics to build two types of models in order to address this issue. One class assumed that proportions close to each other are likely to move in the same average direction whereas the other assumed that the proportions are likely to follow the same pattern. Finally, while it appears to be an unrelated topic, chapter 4 discusses tree-based mixtures of probability simplices. The mixture is tree-based because the simplices share vertices in a manner that can be organized as a tree. Here, the vertex corresponding with the root node is shared by all simplices whereas the vertices corresponding to the leaf nodes belong only to one simplex. However, the ideas used to fit these models might be useful because different clusters might move in different directions and it would be helpful to identify these clusters.

Indeed, now we discuss how these ideas might be united because it provides a useful structure to discuss how these ideas might be expanded upon. These ideas might be particularly useful in modeling high dimensional data that lie on a simplex. In this setting, it might be difficult to impose one global transformation on the data, in part because of the sparsity of the data. Instead, it might be easier to first divide the simplex into a mixture of convex hulls based on how the data is distributed at each time point and then model how these hulls change over time.

Convex hulls are chosen because they enclose an area of a simplex specified by the convex combination of these points. As a result, the area bounded is much more specific. Because these hulls lie in a simplex, the ambient dimension of these hulls is one less than the number of points used to define them. A mixture of convex hulls is used to capture the many pockets of data. Using one hull for each pocket allows the hull to remain as parsimonious as possible.
Further parsimony might be achieved by assuming that some of these points may be shared. One such arrangement is discussed in chapter 4 in which the shared vertices can be organized in a tree. However, these trees require that all paths are the same lengths. If this assumptions are relaxed, it is worth exploring whether such structures are identifiable. We now have to deal with simplices of different dimensions and depending on the distribution, different normalization constants. As it might be discussed in future work by Chakraborty et al., this might give more weight to lower dimensional structures due to the probability being higher. Further, the Hausdorff metric between simplices of different dimensions might be zero so a different metric might be needed. Alternatively, another assumption that might be relaxed is that a simplex is defined by one path and instead by the nodes on the subtree. This would lead to greater parsimony in the model and perhaps allow more information to reach the vertices associated with the leaf nodes. Finally, it might be possible that an entirely different acyclic graph might better represent how the vertices are shared. In all these case, it would be of interest to know when such arrangements are identifiable. With that information, we might develop a model that can toggle between some or all of these different arrangements. These would all be important extensions of the ideas in chapter 4.

Then, we can apply the ideas from the chapters 2 and 3 on the data contained within these convex hulls and model how they change. Obviously, a first step towards using these ideas is ensuring that they work for dimensions higher than two. In the chapter on random directions, we provide background on the higher dimension analogues. These higher dimensional distributions can be plugged into our models in place of their lower dimensional analogues. Meanwhile, the math for the higher dimensional Hellinger geodesic for the Rotational Geodesic Error model has been worked out. However, there might be other geodesics that might be more suitable for the higher cases. Because of this, there still is work implementing these models in higher dimensions.

Next, modeling how the data contained within the convex hulls change depends on whether the data is labeled or not. If the data is labeled, it is possible to track how the data changed from one time point to the next. Then, after the ideas in chapters 2 and 3 have been applied, it might be appropriate to re-update the cluster assignment based how the observations are modeled to change. In other words, information on how the data is changed might help identified how the vertices of convex hulls are shared because clusters might be determined based on which observations move together. This would be an interesting extension of the ideas discussed in the three chapters.

On the other hand, there are two options if the data is not labeled. First, it might make sense to use information based on how the convex hulls and their organization change over time. This could be done by matching convex hulls with their "closest" hull under an appropriate metric. Or, this might be controlled with Bayesian nonparametric priors for random partitions because the hulls should have some probability of splitting, merging with other hulls, or remaining intact while changing [49]. This would be an interesting extension of the ideas in Chapter 4 to understand when
the temporal evolution of structures are identified. An alternative approach is motivated by the original interest in modeling the income proportions from Los Angeles County. HMDA data is collected annually, but only has demographic information on people buying homes with mortgages. On the other hand, US Census data is collected decennially, but contains demographic information on the entire neighborhood. The goal is to be able to infer changes in US Census data from year to year based on the changes in HMDA data and changes from the Census to the HMDA data. It is of interest to find the start and end point when one or both are latent. With such a model, the approach discussed in the previous paragraph might be more appropriate. As a result, one interesting future direction for this scenario is developing a model that can appropriately balance information from the movement and the clusters.

In the development of the ideas discussed in the previous paragraph, it is also important to devise fast, efficient computation schemes. As an example of this importance, the number of articles from the New York Times corpus that could be analyzed was limited by memory constraints. This is problematic because as the models grow increasingly complex, more data is needed to accurately fit these models. However, as the dimension increases, the memory size for a single observation grows as well, potentially restricting the amount of data that can be used for computation.

Fortunately, each chapter has directions to expand on. In chapter 2, the model has to be fitted with different extreme points as its starting point and the results have to be compared. As the dimension grows, there are more extreme points to try and it becomes harder to visualize which transformation is correct. To save this computation, we might use the "means" at each time point to initialize the model. Here, "means" are in quotation because they may be the sample means adjusted based on the general affine transformation and the noise model. Next, because the ideas in chapter 3 heavily rely on Gaussian processes, it might be possible to use the literature on speeding up Gaussian processes [41, 38]. Alternatively, we took advantage of the fact that the Gaussian process is a multivariate normal distribution and used elliptical slice sampling. To better sample from these models or the higher dimensional equivalents, we might combine Multiple Try MCMC with elliptical slice sampling. This could provide a principled approach to consider the entire range of angles during each proposal step for the next mean angles and thus result in better mixing. In any case, it is key that a speed-up is found because after each transformation, the models in chapter 3 might need to be re-learned to account for the difference between the transformed data from one time point and observed data in the next time point. Finally, only the geometric methods provided us multiple fits for the subset of articles in a reasonable amount of time. A natural question worth exploring is how to extend these methods for more general organizations of the tree structure. It was relatively straightforward to apply existing geometric algorithms to learn the simplices defined by the path.

Finally, this framework also presents an opportunity for theory. As stated above, it is of interest
to understand when these structures are identifiable, particularly if they are time varying. In addition, because the data’s dimension might negatively affect the quality of the estimates, it is beneficial to quantify the optimal degradation these estimates suffer. This might be done by calculating the asymptotic rate for the estimates to converge to the true value and examining how quickly the data needs to grow compared to the dimension. Because the model involves finding pockets of data with the convex hull, this asymptotic rate will also be computed according to the number and size of these pockets. It is possible that the algorithm might particularly suffer if there are lots of pockets of thinly spread data. Simulation studies might help to provide some intuition and derive possible rates.

In conclusion, this thesis lays out a solid foundation for modeling how data that lie on a simplex change over time. Within this framework, there remains many other interesting questions in modeling, computation, and theory that are worth exploring. Answering them and uncovering the full potential of this framework will be a tremendous benefit to statistics.
APPENDIX A

Simplex Movement Appendix

A.1 Affine transformation properties

In addition to the properties listed earlier, it is worth wondering what other properties the affine transformation might have. First, the affine transformation and the general affine transformation are not commutative.

Property A.1.1. For \( x, v_i, v_j \in \Delta^D, \eta_i, \eta_j \in (-\infty, 1], v_i \neq v_j, \)

\[
T_{\eta_j, v_j} \circ T_{\eta_i, v_i}(x) \neq T_{\eta_i, v_i} \circ T_{\eta_j, v_j}(x).
\]

Proof. Because it is easier, we will prove this property for the affine transformation. Denote the transformations \( T_{\eta_i, e_i}(x) \) and \( T_{\eta_j, e_j}(x) \) as \( T_i(x) \) and \( T_j(x) \) respectively. Note that

\[
T_j \circ T_i(x) = (1 - \eta_j)((1 - \eta_i)x + \eta_i e_i) + \eta_j e_j
\]

\[
= (1 - \eta_j)(1 - \eta_i)x + (1 - \eta_j)\eta_i e_i + \eta_j e_j
\]

\[
= (1 - \eta_j)(1 - \eta_i)x + \eta_i e_i + \eta_j e_j - \eta_j \eta_i e_i,
\]

and so

\[
T_i \circ T_j(x) = (1 - \eta_j)(1 - \eta_i)x + \eta_i e_i + \eta_j e_j - \eta_j \eta_i e_i.
\]

Because \( \eta_j \eta_i e_j \neq \eta_j \eta_i e_i \) and \( i \neq j \), then \( T_j \circ T_i \neq T_i \circ T_j \). Even if we use the general affine transformation, \( v_i \neq v_j \) so \( \eta_j \eta_i v_j \neq \eta_j \eta_i v_i \).

However, while it is not commutative, both transformations are associative.

Property A.1.2. For \( x, v_i, v_j, v_k \in \Delta^D, \eta_i, \eta_j, \eta_k \in (-\infty, 1], v_i \neq v_j \neq v_j, \)

\[
(T_{\eta_k, v_k} \circ T_{\eta_j, v_j}) \circ T_{\eta_i, v_i}(x) \neq T_{\eta_k, v_k} \circ (T_{\eta_j, v_j} \circ T_{\eta_i, v_i})(x).
\]
We restate the generative description of the additive Dirichlet error model. Given \( \alpha \), \( B \), and \( \eta \), let \( x \) be a coordinate vector. For a given coordinate vector \( x \), \( e \), \( \gamma \), and \( \eta \), let \( T_{\eta, e_1}(x) \) and \( T_{\eta, e_2}(x) \) as \( T_1(x) \), \( T_2(x) \), and \( T_3(x) \) respectively. From the proof of Property A.1.1,

\[
T_j \circ T_i(x) = (1 - \eta_j)(1 - \eta_i)x + (1 - \eta_j)\eta_i e_i + \eta_j e_j,
\]

and

\[
T_k \circ T_j(x) = (1 - \eta_k)(1 - \eta_j)x + (1 - \eta_k)\eta_j e_j + \eta_k e_k.
\]

Then,

\[
T_k \circ (T_j \circ T_i(x)) = (1 - \eta_k)((1 - \eta_j)(1 - \eta_i)x + (1 - \eta_j)\eta_i e_i + \eta_j e_j) + \eta_k e_k = (1 - \eta_k)(1 - \eta_j)(1 - \eta_i)x + (1 - \eta_k)(1 - \eta_j)\eta_i e_i + (1 - \eta_k)\eta_j e_j + \eta_k e_k,
\]

and

\[
(T_k \circ T_j) \circ T_i(x) = (1 - \eta_k)(1 - \eta_j)((1 - \eta_i)x + \eta_i e_i) + (1 - \eta_k)\eta_j e_j + \eta_k e_k = (1 - \eta_k)(1 - \eta_j)(1 - \eta_i)x + (1 - \eta_k)(1 - \eta_j)\eta_i e_i + (1 - \eta_k)\eta_j e_j + \eta_k e_k.
\]

Note that if \( x_l \leq -\frac{\eta_i}{1 - \eta_i} \) for \( l = i, j, \) or \( k \), then we can still write the transformation with the respective \( \eta_i \) equal to \( \frac{\eta_i}{1 - \eta_i} \) and repeat the argument above. Further, if we use \( v_i \), \( v_j \), and \( v_k \) instead of \( e_i \), \( e_j \), and \( e_k \), the argument still holds. Hence, the transformation is associative.

\[ \square \]

### A.2 Additive Dirichlet error model

We restate the generative description of the additive Dirichlet error model. Given \( \alpha \in \mathbb{R}^{(D+1)^+} \) and a coordinate vector \( e_i \), \( i = 1, 2, \ldots, D + 1 \), it can easily be described in a generative manner for a data set, \( x_1, x_2, \ldots, x_N \in \Delta^D \), and parameters, \( \eta, \gamma \in (-\infty, 1] \):

1. For a given coordinate vector \( e_i \) and \( \eta \), let \( z_n = T_{\eta, e_i}(x_n) \) for \( n = 1, 2, \ldots, N \).

2. Draw \( \epsilon_n \sim \text{Dir}(\alpha) \) for \( n = 1, 2, \ldots, N \).

3. For \( \gamma \in [0, 1] \), let \( y_n = (1 - \gamma)z_n + \gamma \epsilon \) for \( n = 1, 2, \ldots, N \).

Then, using change of variables, the likelihood for \( y_n \mid x_n, \gamma, \eta \) is the following:

\[
P(y_n \mid x_n, \gamma, \eta) = \prod_{n=1}^{N} \frac{1}{B(\alpha)} \frac{1}{\gamma^d} \prod_{j=1}^{d} \frac{(y_n)_j - (1 - \gamma)(T_{\eta, e_i}(x_n)_j)}{\gamma}^{\alpha_j - 1}, \tag{A.1}
\]

where \( B(\alpha) = \frac{\prod_{j=1}^{d} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{d} \alpha_j)}. \)
Because it will be easier to work with, we will use the average log likelihood, i.e.

\[
\frac{1}{n} \left( \sum_{n=1}^{N} - \log(B(\alpha)) - \sum_{j=1}^{d} (\alpha_j - 1) (\log((y_n)_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j) - \log(\gamma)) - \log(\gamma) \right).
\]

Using this log likelihood, we can derive the gradients and simplicial constraints for \(\gamma\) and \(\eta\).

### A.2.1 \(\gamma\) log likelihood gradient

**Lemma A.2.1.** Given \(\eta, \alpha\), and a coordinate vector \(e_i\) for some \(i = 1, 2, \ldots, D + 1\), the gradient for \(\gamma\) based on the average log likelihood is:

\[
\frac{d}{d\gamma} \log \left( \sum_{n=1}^{N} \sum_{j=1}^{d} (\alpha_j - 1) \log ((y_n)_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j) \right) - \sum_{j=1}^{d} \frac{\alpha_j}{\gamma}.
\]

**Proof.**

\[
\frac{d}{d\gamma} \log \left( \sum_{n=1}^{N} \sum_{j=1}^{d} (\alpha_j - 1) \log ((y_n)_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j) \right) - \sum_{j=1}^{d} \frac{\alpha_j}{\gamma}.
\]

**Lemma A.2.2.** The simplicial constraints are satisfied if

\[
\gamma \geq \max_{n,j} \left( -\frac{(y_n)_j - T_{\eta,e_i}(x_n)_j}{T_{\eta,e_i}(x_n)_j} \right).
\]
Proof. If \( x, y \in \Delta^d \), then

\[
\sum_j y_j - (1 - \gamma)T_{\eta,e_i}(x)_j = \sum_j y_j - \sum_j (1 - \gamma)T_{\eta,e_i}(x)_j = 1 - (1 - \gamma) = \gamma.
\]

Hence, it suffices to check that \( y_j - (1 - \gamma)T_{\eta,e_i}(x)_j \geq 0 \ \forall j = 1, 2, \ldots, d \). This holds if for an individual point,

\[
\gamma \geq \max_j \left( -\frac{y_j - T_{\eta,e_i}(x)_j}{T_{\eta,e_i}(x)_j} \right),
\]

and for an entire data set,

\[
\gamma \geq \max_n,j \left( -\frac{(y_n)_j - T_{\eta,e_i}(x_n)_j}{T_{\eta,e_i}(x_n)_j} \right).
\]

\( \Box \)

### A.2.2 \( \eta \) log likelihood gradient

**Lemma A.2.3.** Given \( \gamma, \alpha \), and a coordinate vector \( e_i \) for some \( i = 1, 2, \ldots, D + 1 \), the gradient for \( \eta \) based on the average log likelihood is:

\[
\frac{d}{d\eta} \log \mathbb{E}_{y \mid x, \gamma, \eta} = \frac{1}{n} \left( \sum_{n=1}^N \sum_{j=1}^d \frac{(\alpha_j - 1)(1 - \gamma)((x_n)_j - (e_i)_j)}{y_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j} \mathbb{I}(\xi_n = 1) \right).
\]

**Proof.** For active \( x_n \)'s and their \( y_n \)'s

\[
\frac{\partial}{\partial \eta} \log \mathbb{E}_{y \mid x, \gamma, \eta} = \frac{1}{n} \left( \sum_{n=1}^N \left( \sum_{j=1}^d (\alpha_j - 1)(\log((y_n)_j - (1 - \gamma)T_{\eta,e_i}(x_n)_j)) - \log(\gamma) \right) - d \log(\gamma) \right)
\]

\[
= \frac{1}{n} \sum_{n=1}^N \sum_{j=1}^d \frac{(\alpha_j - 1)(1 - \gamma)((x_n)_j - (e_i)_j)}{(y_n)_j - (1 - \gamma)T_{\eta,e_i}(x)_j}.
\]

\( \Box \)
Lemma A.2.4. For active $x_n$, the simplicial constraint is satisfied if

$$\max_{j \neq i} \left( \frac{(y_n)_j - (1 - \gamma)(x_n)_j}{(1 - \gamma)(x_n)_j} \right) \leq \eta \leq \left( \frac{(y_n)_i - (1 - \gamma)(x_n)_i}{(1 - \gamma)(1 - (x_n)_i)} \right).$$

Proof. From the simplicial constraint, we get that

$$0 \leq y - (1 - \gamma)((1 - \eta)x_n + \eta e_i)$$

$$= y - (1 - \gamma)(x_n - \eta(x_n - e_i))$$

$$= y - (1 - \gamma)x_n + \eta(1 - \gamma)(x_n - e_i).$$

So, as

$$-(y - (1 - \gamma)x_n) \leq \eta(1 - \gamma)(x_n - e_i),$$

and $(x_n)_i - 1 \leq 0$, then

$$\left( \frac{-(y)_j - (1 - \gamma)(x_n)_j}{(1 - \gamma)(x_n)_j} \right) \leq \eta \leq \left( \frac{(y)_i - (1 - \gamma)(x_n)_i}{(1 - \gamma)(1 - (x_n)_i)} \right).$$

Because this must hold $\forall j$, we take the appropriate max’s to get the lemma.

Lemma A.2.5. For inactive $x_n$ and its corresponding $y_n$ and $\eta \leq 0$, if

$$\max_{n,j \neq 1} \left( \frac{T_{\eta_1}(x_n)_j}{y_n} \right) \leq \frac{1}{1 - \gamma},$$

then the simplicial constraint is satisfied.

Proof. A simple rearrangement of the simplicial constraints gives us $\forall j = 1, 2, \ldots, d$,

$$\frac{T_{\eta_1}(x_n)_j}{(y_n)_j} \leq \frac{1}{1 - \gamma}.$$

On an inactive set, $\frac{T_{\eta_1}(x_n)_i}{(y_n)_i}$ must be 0. Hence,

$$\max_{n,j \neq i} \left( \frac{T_{\eta_1}(x_n)_j}{y_n} \right) \leq \frac{1}{1 - \gamma}$$

must be true or else, the log likelihood is undefined.
A.2.3 Multiple direction affine transformation with additive Dirichlet error

As shown above, we can derive the gradients for $\gamma$ and $\eta$ in the case of one direction. However, it might be that one direction is not rich enough. One way to make the model richer is by using affine transformations in multiple directions. In other words, given $\alpha \in \mathbb{R}^{(D+1)+}$, our generative model becomes the following for a data set, $x_1, x_2, \ldots, x_N \in \Delta^D$, and parameters, $\eta_1, \eta_2, \ldots, \eta_{D+1} \in (-\infty, 1]$, $\gamma \in (0, 1)$:

1. For $i \in 0, 1, 2, \ldots, D+1$, a given coordinate vector $e_i$, and $\eta_i$, let $z_i^{n+1} = T_{\eta_i e_i}((z_i^n)_n)$. Here $z_0^n = x_n$.
2. Draw $\epsilon \sim \text{Dir}(\alpha)$.
3. Let $y_n = (1 - \gamma)z^{D+2} + \gamma \epsilon$.

If we introduce some other notation, we can also write out the transformation explicitly.

- Define $T_{a:b}(x) := T_{\eta_b e_b} \circ T_{\eta_{b-1} e_{b-1}} \circ \cdots \circ T_{\eta_a e_a}(x)$ if $b > a$ and $x$ otherwise.
- Define $T_{-\eta}(x) := T_{\eta_{i+1}:d}(T_{1:i-1}(x))$.
- Define $T_{-\eta_i}(x) := T_{\eta_i} - \prod_{j=i+1}^d (1 - \beta_j)e_j$.
- Define $\tilde{\eta}_i = \max(\eta_i, \frac{-\epsilon_i}{1 - \epsilon_i})$. Say $\eta_i$ is active for $x_n$ if $\tilde{\eta}_i = \eta_i$ and inactive otherwise. If necessary, we will indicate with $\xi^i_n$. Again, $\xi^i_n = 1$ if $\eta_i$ is active for $x_n$ active and $\xi^i_n = 0$ otherwise.

Using this notation, our transformation is the following.

$$T_{1:d}(x_n) = \prod_{i=1}^d (1 - \tilde{\eta}_i)x_n + \sum_{i=1}^d \prod_{j=i+1}^d (1 - \tilde{\eta}_j)\tilde{\eta}_i e_i.$$  

As a result, using change of variables and modifying the likelihood from before, we get that the likelihood is

$$P(y_n | x_n, \gamma, \eta_1, \eta_2, \ldots, \eta_{D+1}) = \prod_{n=1}^N \frac{1}{B(\alpha)} \frac{1}{\gamma^d} \prod_{j=1}^d \left( \frac{(y_n)_j - (1 - \gamma)T_{1:d}(x_n)_j}{\gamma} \right)^{\alpha_j - 1},$$

where $B(\alpha) = \frac{\prod_{i=1}^d \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^d \alpha_j)}$. The log likelihood is

$$\frac{1}{n} \left( \sum_{n=1}^N \log\left( \frac{1}{B(\alpha)} \right) - \sum_{j=1}^d (\alpha_j - 1) \log((y_n)_j - (1 - \gamma)T_{1:d}(x_n)_j) - \log(\gamma) - \log(\gamma) \right).$$
By writing out the log likelihood, we see that we can still use Lemma 2.3.1 and Lemma 2.3.2 for \( \gamma \) if we use \( T_{1:d}(x) \) in place of \( T_{e_i}(x) \). As such, the difficulty in fitting such a model is with \( \eta_i \). In particular, what makes it difficult to estimate \( \eta_i \) is that \( \eta_i \) can affect \( \eta_j, i < j \leq D + 1 \).

First, as the following lemma shows, \( \eta_i \) can affect whether future \( \eta_j \)'s are active or not.

**Lemma A.2.6.** For \( i < j \leq d \) and given \( \eta_i \),

\[
\eta_i \leq 1 + \frac{\tilde{\eta}_j}{\prod_{k=i+1}^{j} (1 - \tilde{\eta}_k)(z^i_n)_j},
\]

\( \eta_j \) is active for \( x_n \).

**Proof.** First, we tackle the case of \( j = i + 1 \). By definition, \( \eta_{i+1} \) for \( x_n \) is active if

\[
(1 - \eta_i)((z^i_n)_n)_{i+1} \geq \frac{-\eta_{i+1}}{1 - \eta_{i+1}}
\]

\[
(1 - \eta_i) \geq \frac{-\eta_{i+1}}{(1 - \eta_{i+1})((z^i_n)_n)_{i+1}}
\]

\[\eta_i \leq 1 + \frac{\eta_{i+1}}{(1 - \eta_{i+1})((z^i_n)_n)_{i+1}}.\]

Then, by definition, \( \eta_j, j > i + 1 \), is active if

\[
\prod_{k=i+1}^{j-1} (1 - \tilde{\eta}_k)(1 - \eta_i)(z^i_n)_j \geq \frac{-\eta_j}{1 - \eta_j}
\]

\[
(1 - \eta_i) \geq \frac{-\eta_j}{\left( \prod_{k=i+1}^{j} (1 - \tilde{\eta}_k) \right)(z^i_n)_j}
\]

\[\eta_i \leq 1 + \frac{\eta_j}{\left( \prod_{k=i+1}^{j} (1 - \tilde{\eta}_k) \right)(z^i_n)_j}.
\]

The corollary below immediately follows from the lemma above, which tells us for what values of \( \eta_i \) all subsequent \( \eta_j \)'s will be active.

**Corollary A.2.1.** For \( i < k \leq j \leq D + 1 \), \( \eta_k \) is active for \( x_n \) iff

\[
\eta_i \leq 1 + \frac{\eta_j}{\prod_{k=i+1}^{j} (1 - \eta_k)(z^i_n)_j}
\]
In addition to affecting if $\eta_j$ is active, the following lemma shows that $\eta_i$ can affect the value of $\eta_j$ if $\eta_j$ is inactive.

**Lemma A.2.7.** If $\eta_j$ is inactive for $i < j \leq D + 1$ for $x_n$, then

$$
\tilde{\eta}_j = \frac{-j - 1 \prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)(z_n^i)_j}{1 - \prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)(z_n^i)_j}.
$$

**Proof.** If $\eta_j$ is inactive for $x_n$ for $i < j \leq D + 1$, then

$$
\tilde{\eta}_j = \frac{-z_j^j}{1 - (z_n^j)_j} = \frac{-(1 - \tilde{\eta}_{j-1})z^{j-1} + \tilde{\eta}_{j-1}e_{j-1}}{1 - ((1 - \tilde{\eta}_{j-1})z^{j-1} + \tilde{\eta}_{j-1}e_{j-1})} = \frac{-z_j^{j-1}}{1 - (1 - \tilde{\eta}_{j-1})z_{j-1}^{j-1}} = \frac{-(1 - \tilde{\eta}_{j-2})(1 - \tilde{\eta}_{j-1})z_{j-2}^{j-2}}{1 - (1 - \tilde{\eta}_{j-2})(1 - \tilde{\eta}_{j-1})z_{j-2}^{j-2}}$

$$
\vdots
$$

$$
= \frac{- \prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)(z_n^i)_j}{1 - \prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)(z_n^i)_j}.
$$

As a result, the log likelihood gradients with respect to $\eta_i$ changes depending on whether all subsequent $\eta_j$’s are active or not. As a point of comparison, the following lemma gives us the gradient if all subsequent $\eta_j$’s are active.

**Lemma A.2.8.** For $\eta_i$, if all subsequent $\eta_j$’s are active for $x_n$, the partial derivative is

$$
\sum_{j=1}^{d} \frac{(\alpha_j - 1)(1 - \gamma)(T_{-\eta_i}(x_n))_j}{(y_n)_j - (1 - \gamma)T_{1:D+1}(x_n)}.
$$
Proof. Using notation from earlier,

\[
\frac{d}{d\eta_i} T_{1:D+1}(x_n) = \frac{d}{d\eta_i} \prod_{j=1}^d (1 - \tilde{\eta}_j)x_n + \sum_{j=1}^d \prod_{k=i+1}^d (1 - \tilde{\eta}_k)\tilde{\eta}_j e_j
\]

\[
= -\prod_{j=1, j\neq i}^d (1 - \tilde{\eta}_j)x_n - \sum_{j=1, k=j+1, j\neq i}^d \prod_{k=i+1}^d (1 - \tilde{\eta}_k)\tilde{\eta}_j e_j + \prod_{k=i+1}^d (1 - \tilde{\eta}_k)e_i
\]

\[
= -(T_{\eta_i}^-(x_n)).
\]

The rest follows from the proof of Lemma 2.3.3. \[\square\]

On the other hand, as the following lemma shows, we need to use back-propagation to calculate the derivative of \(\eta_j\) if there are any subsequent, inactive \(\eta_j\).

Lemma A.2.9. Using notation from before, for \(\eta_{i+1}\) inactive for \(x_n\)

\[
\frac{d}{d\eta_i} \tilde{\eta}_{i+1} = \frac{(z^i_{n})_{i+1}}{(1 - \eta_i) - (z^i_{n})_{i+1}} \frac{1}{(1 - \eta_i)^2}.
\]

and for \(\eta_j\) inactive for general \(x_n\),

\[
\frac{d}{d\eta_i} \tilde{\eta}_j = \frac{(z^i_{n})_j}{\prod_{k=i}^j 1 - \eta_k} \frac{1}{(\prod_{k=i}^j 1 - \eta_k)^2} \frac{1}{(\prod_{k=i}^j 1 - \eta_k)^2} \left( \prod_{k=i+1}^{j-1} (1 - \eta_k) + \sum_{i\leq l\leq j-1} \mathbb{I}(\zeta^i_{n} = 1) \frac{d}{d\eta_l} \tilde{\eta}_l \prod_{k=i, k\neq l}^d (1 - \eta_k) \right).
\]

Proof. For \(\eta_{i+1}\) inactive for \(x_n\),

\[
\frac{d}{d\eta_i} \tilde{\eta}_{i+1} = \frac{d}{d\eta_i} \frac{(1 - \eta_i)(z^i_{n})_{i+1}}{1 - (1 - \eta_i)(z^i_{n})_{i+1}}
\]

\[
= -\frac{d}{d\eta_i} \frac{(z^i_{n})_{i+1}}{\prod_{k=i}^j 1 - \eta_k} \frac{1}{(1 - \eta_i) - (z^i_{n})_{i+1}} \frac{1}{(1 - \eta_i)^2}.
\]
For general \( \eta_j, j > i + 1 \), we can repeat the argument above and get that

\[
\frac{d}{d\eta_i} \tilde{\eta}_j = \frac{(z_i^n)^j}{\prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)^2} \frac{1}{\prod_{k=i}^{j-1} (1 - \tilde{\eta}_k)^2} \frac{d}{d\eta_i} \left( \prod_{k=i}^{j-1} 1 - \tilde{\eta}_k \right).
\]

For \( k = i \), the derivative is \( \prod_{k=i+1}^{j-1} 1 - \tilde{\eta}_k \). For \( k \neq i \), the derivative is zero unless \( \tilde{\eta}_k \) is inactive for \( x_n \). As a result,

\[
\frac{d}{d\eta_i} \left( \prod_{k=i}^{j-1} 1 - \tilde{\eta}_k \right) = \left( \prod_{k=i+1}^{j-1} (1 - \tilde{\eta}_k) + \sum_{i \leq l \leq j-1} \mathbb{I} (\xi^l_n = 1) \frac{d}{d\eta_i} \tilde{\eta}_l \prod_{k=i, k \neq l}^{j-1} (1 - \tilde{\eta}_k)^2 \right)\left( \prod_{k=i}^{j-1} 1 - \tilde{\eta}_k \right)\left( (y_n)^j - (1 - \gamma)T_{1:D+1}(x_n) \right).
\]

Then, we can use the following lemma to adjust the derivative for \( \eta_i \) if there are any inactive subsequent \( \eta_j \).

**Lemma A.2.10.** For \( \eta_i \), if some subsequent \( \eta_j \)'s are inactive for \( x_n \), the partial derivative is

\[
\sum_{j=1}^{d} \frac{\alpha_j - 1}{(1 - \gamma)} \left( T_{-\eta_i}(x) + \sum_{l=1}^{d} \mathbb{I} (\xi^l_n = 1) T_{-\eta_i}(x) e_{j} \right) (y_n)^j - (1 - \gamma)T_{1:D+1}(x_n) \frac{d\eta_i}{d\eta_i}.
\]

where \( \frac{d\eta_i}{d\eta_i} \) is given by Lemma A.2.9.

**Proof.** By the product rule,

\[
\frac{d}{d\eta_i} T_{1:D+1}(x_n) = \frac{d}{d\eta_i} \prod_{j=1}^{d} (1 - \tilde{\eta}_j) x_n + \sum_{j=1}^{d} \prod_{k=i+1}^{d} (1 - \tilde{\eta}_k) e_j
\]

\[
= - \prod_{j=1, j \neq i}^{d} (1 - \tilde{\eta}_j) x_n - \sum_{j=1, k=j+1}^{d} \prod_{k \neq i}^{d} (1 - \tilde{\eta}_k) e_j + \prod_{k=i+1}^{d} (1 - \tilde{\eta}_k) e_i +
\]

\[
\sum_{l=1}^{d} \mathbb{I} (\xi^l_n = 0) \sum_{j=1}^{d} \prod_{j=1, j \neq l}^{d} (1 - \tilde{\eta}_j) e_j + \prod_{k=l+1}^{d} (1 - \tilde{\eta}_k) e_l \frac{d\eta_i}{d\eta_i}.
\]

\[
= - \left( T_{-\eta_i}(x_n) - \sum_{l=1}^{d} \mathbb{I} (\xi^l_n = 1) T_{-\eta_i}(x_n) \frac{d\eta_i}{d\eta_i} \right).
\]
The rest follows from the proof of Lemma 2.3.3.

Now that we have the gradients, we again have to check if the simplicial constraints are satisfied. Suppose we take the approach as before and partition \( \eta_i \) into the intervals \((b_i^{(1)}, b_i^{(2)}, \ldots, b_i^{(n)})\). If \( x_n \) is active for all subsequent \( \eta_j \)'s, we just have a simple refinement, as shown in the following lemma.

**Lemma A.2.11.** For \( \eta_i \) with \( \eta_j \), \( 1 \leq j \leq D + 1 \), fixed and \( \xi_k^i = 1 \) for \( i \leq k \leq D + 1 \), the simplicial constraints are satisfied if

1. for \( T_{\eta_i}(x_n)_l \neq 0, 1 \leq l \leq D + 1 \),

\[
\max_{T_{\eta_i}(x_n)_l > 0} \frac{-(y)_l - (1 - \gamma)T_{-\eta_i}(x_n)_l}{(1 - \gamma)T_{-\eta_i}(x_n)_l} \leq \eta_i \leq \min_{T_{\eta_i}(x_n)_l < 0} \frac{-(y)_l - (1 - \gamma)T_{-\eta_i}(x_n)_l}{(1 - \gamma)T_{-\eta_i}(x_n)_l},
\]

2. for \( T_{\eta_i}(x_n)_l = 0, 1 \leq l \leq D + 1 \),

\[
y - (1 - \gamma)(T_{-\eta_i}(x_n))_l \geq 0.
\]

**Proof.**

\[
0 \leq y - (1 - \gamma)T_{\eta_i}(x_n)
= y - (1 - \gamma) \left( \prod_{j=1}^{d} (1 - \eta_j) x_n + \sum_{j=1}^{d} \prod_{k=j+1}^{d} (1 - \eta_k) \tilde{\eta}_j e_j \right)
= y - (1 - \gamma) \left( \prod_{j=1, j \neq i}^{d} (1 - \eta_j) x_n - \eta_i \prod_{j=1, j \neq i}^{d} (1 - \tilde{\eta}_j) x_n + \sum_{j=1, k=j+1, k \neq i}^{d-1} \prod_{k=j+1}^{d} (1 - \tilde{\eta}_k) \tilde{\eta}_j e_j - \eta_i \sum_{j=1, k=j+1, k \neq i}^{d-1} \prod_{k=j+1}^{d} (1 - \tilde{\eta}_k) \tilde{\eta}_j e_j + \eta_i \prod_{k=i+1}^{d} (1 - \tilde{\eta}_k) e_i \right)
= y - (1 - \gamma) \left( T_{-\eta_i}(x_n) - \eta_i \prod_{j=1, j \neq i}^{d} (1 - \eta_j) x_n + \sum_{j=1, k=j+1, k \neq i}^{i-1} \prod_{k=j+1}^{d} (1 - \tilde{\eta}_k) \tilde{\eta}_j e_j - \prod_{k=i+1}^{d} (1 - \tilde{\eta}_k) e_i \right)
= y - (1 - \gamma) \left( T_{-\eta_i}(x_n) - \eta_i T_{-\eta_i}(x_n) \right).
\]
Because this must hold $\forall l \in 1, 2, \ldots, D+1$, then we must check the conditions listed in the lemma.

Unfortunately, we don’t have a way to check for simplicial constraints if there are inactive $\eta$. Lemma A.2.7 makes it difficult to imagine that there is a close formed solution. An alternative approach might be to prove some monotonicity properties of the multiple direction affine transformation. Then, if we are searching over $(b_i^l, b_i^{l+1})$, we might be able to use these properties to check if the minimum is less than 0. Unfortunately, Lemma A.2.15 show us that such an approach might not work.

However, before we state it, we need to establish some properties about whether $x_n$ will be active or not and how $\tilde{\eta}_j$ changes as $\eta_i$ increases. These lemmas will also help us prove Lemma A.2.15.

**Lemma A.2.12.** Given $x_n$, as $\eta_i$ increases and all other $\eta_j$, $1 \leq j \leq D + 1$, are fixed,

- $\eta_{i+1}$ can only change from being active to inactive for $x_n$, and
- if $\eta_{i+1}$ becomes inactive for $x_n$, $\tilde{\eta}_{i+1}$ monotonically increases.

**Proof.** If $\eta_{i+1} \geq 0$, then $\eta_{i+1}$ will always be active for $x_n$. After all, for $\eta_{i+1} \geq 0$, $1 - \eta_{i+1}$ will be non-negative, but $-\eta_{i+1}$ will be negative. As a result, $\tilde{\eta}_{i+1}$ will always be $\eta$ by definition because the following is trivially true:

$$
(z_n^{i+1})_{i+1} \geq 0 \geq -\frac{\eta_{i+1}}{1 - \eta_{i+1}}.
$$

On the other hand, $\tilde{\eta}_{i+1}$ is defined to be the max($\eta_{i+1}$, $-\frac{(z_n^{i+1})_{i+1}}{1 - (z_n^{i+1})_{i+1}}$). Following from our calculation for Lemma A.2.7,

$$
-\frac{(z_n^{i+1})_{i+1}}{1 - (z_n^{i+1})_{i+1}} = -\frac{(1 - \eta_i)(z_n^i)_{i+1}}{1 - (1 - \eta_i)(z_n^i)_{i+1}} = -\frac{(z_n^i)_{i+1}}{1 - \eta_i - (z_n^i)_{i+1}}.
$$

Because $\eta_i$ is increasing, $1 - \eta_i$ must be decreasing monotonically. As a result, $\frac{1}{1 - \eta_i} - (z_n^i)_{i+1}$ must be increasing monotonically. Hence,

$$
-\frac{(z_n^i)_{i+1}}{1 - \eta_i - (z_n^i)_{i+1}}.
$$

must be increasing monotonically. Thus, if $\eta_{i+1} < 0$, $\tilde{\eta}_{i+1} \geq \eta_{i+1}$ for some $c \in (\infty, 1]$. This also shows that for $\eta_i \geq c$, $\tilde{\eta}_{i+1}$ must increase monotonically. \qed
We can generalize this result for subsequent \( j \).

**Lemma A.2.13.** For \( x_n \), as \( \eta_i \) increases and all other \( \eta_k \), \( 1 \leq k \leq D + 1 \), are fixed,

- \( \eta_j \) can only change from being inactive to active for \( x_n \).

- If \( \xi_j \) changes from 1 to 0, \( \eta_{i+1} \) monotonically increases.

**Proof.** If \( \eta_j \) is positive, then \( \eta_j \) will be always active because of the argument given in Lemma A.2.12.

On the other hand, suppose \( \eta_{i+2} \) is negative. Then, by Lemma A.2.7,

\[
\frac{(z_n^{i+2})_{i+2}}{1 - (z_n^{i+2})_{i+2}} = \frac{\prod_{k=i}^{i+1} (1 - \eta_k) (z_n^i)_{i+2}}{\prod_{k=i}^{i+1} (1 - \eta_k) (z_n^i)_{i+2}} = \frac{(z_n^i)_{i+2}}{\prod_{k=i}^{i+1} (1 - \eta_k) - (z_n^i)_{i+2}}.
\]

By Lemma A.2.12, \( \eta_{i+1} \) must be constant or monotonically increasing.

If \( \eta_{i+1} \) is constant for \( \eta_i \in (-\infty, c) \), \( c \in (-\infty, 1] \), \( 1 - \eta_{i+1} \) must be positive by definition. As a result, we can repeat the argument used in Lemma A.2.12 to show that \( \eta_{i+1} \) is increasing \((-\infty, c)\).

If \( \eta_{i+1} \) is monotonically increasing, then \( \prod_{k=i}^{i+1} (1 - \eta_k) \) is the product of two positive, monotonically decreasing functions. Because it’s still monotonically decreasing, \( \frac{1}{\prod_{k=i}^{i+1} (1 - \eta_k)} \) must be monotonically increasing. So, as before, \( \eta_{i+2} \) must be monotonically increasing.

We can then repeat this argument for \( i + 3 \) to show that \( \eta_{i+3} \) must be monotonically increasing, then for \( i + 4 \) to show that \( \eta_{i+4} \) must be monotonically increasing, and so on. \( \square \)

These lemmas do allow us to establish the following.

**Lemma A.2.14.** Assume all other \( \eta_j \), \( 1 \leq k \leq D + 1 \) and \( k \neq i \), are fixed. Then, for \( T_{1:D+1}(x_n)_k \), \( 1 \leq j \leq D + 1 \) and \( j \neq i \),

- If \( \eta_j \) is inactive for \( x_n \), \( T_{1:D+1}(x_n)_j = 0 \).

- If \( \eta_k \) is active for \( x_n \), \( T_{1:D+1}(x_n)_j \) monotonically decreases as a function of \( \eta_i \).

**Proof.** If \( \eta_j \) is inactive for \( x_n \), because \( (e_k)_j = 0 \) for \( k \neq j \), \( T_{1:D+1}(x_n)_j = 0 \).
Now assume that \( \eta_k \) is active for \( x_n \). For \( j < i \), by similar logic as above,

\[
(T_{1:D+1}(x))_k = (T_{j<k\leq D+1}(z_n^{k+1}))_k \\
= \left( \prod_{k=k+1}^{i-1} (1 - \tilde{\eta}_k) \right) \prod_{k=i}^{D+1} (1 - \tilde{\eta}_k).
\]

By Lemma A.2.12 and A.2.13, \( (1 - \tilde{\eta}_k) \) for \( i \leq k \leq D + 1 \) is either a strictly decreasing or constant function as \( \eta_i \) increases. As all are positive and strictly decrease from their constant value, \( \prod_{k=i}^D (1 - \tilde{\eta}_k) \) must be a strictly decreasing function. Hence, for \( j < i \), \( (T_{1:D+1}(x))_j \) is a decreasing function.

For \( k > i \), because \( \eta_i \) can vary, we have to start our argument from the \( i \).

\[
(T_\eta(x))_j = (T_{i:D+1}((z_n^i)))_j \\
= \left( \prod_{k=i}^{D+1} (1 - \tilde{\eta}_k)(z_n^i)_j + \sum_{k=i}^{D+1} \prod_{l=k+1}^{D+1} (1 - \tilde{\eta}_l)\eta_k e_k \right) j \\
= \left( \prod_{k=i}^{D+1} (1 - \tilde{\eta}_k)(z_n^i)_j + \prod_{l=j+1}^{D+1} (1 - \tilde{\eta}_l)\eta_j \right) \\
= \left( \prod_{k=j+1}^{D+1} (1 - \tilde{\eta}_k) \right) \left( \prod_{k=i}^{j} (1 - \tilde{\eta}_k)(z_n^i)_j - \eta_j \right).
\]

Note that because \( \xi_n^j = 1 \),

\[
\left( \prod_{k=i}^{j} (1 - \tilde{\eta}_k)(z_n^i)_j - \eta_j \right) = (1 - \eta_j)(z_n^i)_j - \eta_j \geq 0.
\]

Again, by Lemma A.2.12 and A.2.13,

\[
\left( \prod_{k=j+1}^{d} (1 - \tilde{\eta}_k) \right),
\]

and

\[
\left( \prod_{k=i}^{j} (1 - \tilde{\eta}_k)(z_n^i)_j - \eta_j \right),
\]

are strictly decreasing, positive function as \( \eta_i \) increases. As a result, then for \( j > i \), \( (T_\eta(x_n))_j \) is a decreasing function.

This lemma is encouraging because if we are searching through an interval \( (b_i, b_{i+1}) \), we just
need to find the largest value $b_t$ on the interval such that $T_{1:D+1}(x_n)_j \geq 0$, $j \neq i$. Then, the simplicial constraint is satisfied for $b_t$, $b_t$. Unfortunately, this makes Lemma A.2.15 even more dispiriting.

**Lemma A.2.15.** Assume all other $\eta_j$, $1 \leq k \leq D + 1$ and $k \neq i$, are fixed. $T_{1:D+1}(x)_i$ does not necessarily monotonically increase as a function of $\eta_i$.

**Proof.** For $i$, we have to start our argument from $i$.

\[
(T_{1:D+1}(x))_i = (T_{i:D+1}(z_n^i))_i
\]

\[
= \left( \prod_{k=i}^d (1 - \tilde{\eta}_k)(z_n^i)_i + \prod_{k=i+1}^d (1 - \tilde{\eta}_k)\eta_i \right)
\]

\[
= \prod_{k=i+1}^d (1 - \tilde{\eta}_k) \left( (1 - \eta_i)(z_n^i)_i + \eta_i \right).
\]

By Lemma A.2.12 and A.2.13, $\prod_{k=i+1}^d (1 - \tilde{\eta}_k)$ is a positive decreasing function, but $\left( (1 - \eta_i)(z_n^i)_i + \eta_i \right)$ is a positive increasing function. The product is thus not necessarily monotonically increasing or decreasing.

Interestingly enough, we do have the following corollary because of Lemma A.2.14.

**Corollary A.2.2.** If the simplicial constraints are satisfied, then $T_{1:D+1}(x_n)_i$ monotonically increases as a function of $\eta_i$.

**Proof.** Because

\[
\sum_{j=1}^d T_{1:D+1}(x)_j = 1
\]

\[
T_{1:D+1}(x)_i = 1 - \sum_{j=1, j \neq i}^d T_{1:D+1}(x)_j,
\]

and $T_{1:D+1}(x)_j$ for $1 \leq j \leq D + 1$ and $j \neq i$ decreases by Lemma A.2.14, $T_{1:D+1}(x)_i$ must be monotonically increasing.

However, we are hoping to use monotonicity properties to check the simplicial constraint and not the other way around.

Obviously, switching to using the general affine transformation does not help and would make it even more difficult to estimate $\eta_i$. In addition, these calculations makes it hard to imagine how we
might we recover the transformation if we had multiple additive Dirichlet error in place of or in addition to multiple affine transformation.

Still, even though we couldn’t prove the monotonicity results, this does not mean that the multiple direction affine transformation is unusable. If all \( \eta_i \in [0, 1] \), then \( x_n \) will always be active for all \( \eta_i \). We can apply the approach that we used in the single affine transformation with Lemma A.2.11 and Lemma A.2.8 to derive a MLE estimate for \( \eta_i \) given \( \gamma \). Unfortunately, this might be too limiting because it is fairly straightforward to apply the multiple affine transformation with some \( \eta_i < 0 \).

A.3 Rotational geodesic error model

A.3.1 Geodesics in \( \Delta^2 \)

For ease of presentation, we drop the subscript \( n \).

**Lemma A.3.1.** If \( z \) and \( y \) are our two points on the simplex Hellinger distance geodesic, then

\[
\phi_0 = \tan^{-1} \left( \frac{z_3 y_1 - z_1 y_3}{z_2 y_3 - z_3 y_2} \right),
\]

\[
a = \frac{z_3}{z_1 \cos(\phi_0) + z_2 \sin(\phi_0)}.\]

**Proof.** Note that

\[
\cot(\theta) = a \cos(\phi - \phi_0)
\]

\[
\frac{\cos(\theta)}{\sin(\theta)} = a \left( \cos(\phi) \cos(\phi_0) + \sin(\phi) \sin(\phi_0) \right)
\]

\[
\cos(\theta) = a \left( \sin(\theta) \cos(\phi) \cos(\phi_0) + \sin(\theta) \sin(\phi) \sin(\phi_0) \right)
\]

\[
z_3 = a \left( z_1 \cos(\phi_0) + z_2 \sin(\phi_0) \right).
\]

As a result,

\[
\frac{z_3}{y_3} = \frac{z_1 \cos(\phi_0) + z_2 \sin(\phi_0)}{y_1 \cos(\phi_0) + y_2 \sin(\phi_0)}
\]

\[
\frac{z_3}{y_3} = \frac{z_1 + z_2 \tan(\phi_0)}{y_1 + y_2 \tan(\phi_0)}
\]

\[
z_3 \left( y_1 + y_2 \tan(\phi_0) \right) = y_3 \left( z_1 + z_2 \tan(\phi_0) \right)
\]

\[
z_3 y_1 - z_1 y_3 = \tan(\phi_0) \left( z_2 y_3 - z_3 y_2 \right)
\]

\[
\tan(\phi_0) = \frac{z_3 y_1 - z_1 y_3}{z_2 y_3 - z_3 y_2}
\]

\[
\phi_0 = \tan^{-1} \left( \frac{z_3 y_1 - z_1 y_3}{z_2 y_3 - z_3 y_2} \right).
\]
Then, with $\phi_0$ and the result above,

$$a = \frac{z_3}{z_1 \cos(\phi_0) + z_2 \sin(\phi_0)}$$

if $\cos(\phi - \phi_0) \neq 0$ for $z$. Otherwise,

$$a = \frac{y_3}{y_1 \cos(\phi_0) + y_2 \sin(\phi_0)}.$$ 

Lemma A.3.2. If $\phi_z = \phi_y$ and $z \neq y$, then the geodesic endpoint is $(0, 0, 1)$ if $\theta_z > \theta_y$ and $(\cos(\phi_z), \sin(\phi_z), 0)$ if $\theta_z < \theta_y$.

Proof. Our geodesic equation does not work in the standard spherical coordinates if $\phi_z = \phi_y$ because we get a contradiction. After all, this would imply that

$$\cot(\theta_z) = a \cos(\phi_z - \phi_0) = a \cos(\phi_z - \phi_0) = \cot(\theta_y),$$

but as $z \neq y$, $\cot(\theta_z) \neq \cot(\theta_y)$. However, we can first transform our data and then use the standard spherical coordinates. The derivation of the geodesic only depends on our use of the spherical coordinates and not on how we define the transformation from spherical to Cartesian coordinates. The simplest transformation is to swap $z_1$ and $z_3$ and $y_1$ and $y_3$, i.e.

$$
\begin{pmatrix}
\sin \theta_z \cos \phi_z \\
\sin \theta_z \sin \phi_z \\
\cos \theta_z
\end{pmatrix}
\mapsto
\begin{pmatrix}
\cos \theta_z \\
\sin \theta_z \sin \phi_z \\
\sin \theta_z \cos \phi_z
\end{pmatrix},
\begin{pmatrix}
\sin \theta_y \cos \phi_y \\
\sin \theta_y \sin \phi_y \\
\cos \theta_y
\end{pmatrix}
\mapsto
\begin{pmatrix}
\cos \theta_y \\
\sin \theta_y \sin \phi_y \\
\sin \theta_y \cos \phi_y
\end{pmatrix}.
$$

Then, by Lemma 2.3.6, we have that

$$\phi_0 = \tan^{-1} \left( \frac{z_3 y_1 - z_1 y_3}{z_2 y_3 - z_3 y_2} \right),$$

$$= \tan^{-1} \left( \frac{\sin \theta_z \cos \phi_z \cos \theta_y - \cos \theta_z \sin \theta_y \cos \phi_y}{\sin \theta_z \sin \phi_z \sin \theta_y \cos \phi_y - \sin \theta_z \cos \phi_z \sin \theta_y \sin \phi_y} \right),$$

$$= \tan^{-1} \left( \frac{\cos \phi_z (\sin \theta_z \cos \theta_y - \cos \theta_z \sin \theta_y)}{0} \right),$$

$$= \tan^{-1} \left( \frac{\cos \phi_z \sin(\theta_z - \theta_y)}{0} \right).$$

121
Assuming $\phi_z \neq 0, \frac{\pi}{2}$, then let us first look at $\theta_z < \theta_y$. Then, because $\theta_z, \theta_y \in [0, \frac{\pi}{2}]$,

$$\sin(\theta_z - \theta_y) < 0 \implies \phi_0 = -\frac{\pi}{2}.$$ 

By Lemma 2.3.6, then

$$a = \frac{\sin \theta_z \cos \phi_z}{\cos \theta_z \cos \phi_0 + \sin \theta_z \sin \phi_z \sin \phi_0} = \frac{\sin \theta_z \sin \phi_z \sin \phi_0}{\sin \phi_z} = \frac{\cos \phi_z}{\sin \phi_z} = -\cot \phi_z.$$ 

If we set $\phi' = \frac{\pi}{2}$ and plug in $a$ and $\phi_0$, then

$$\cot \theta' = -\cot \phi_z \cos \left(\frac{\pi}{2} + \frac{\pi}{2}\right) \iff \theta' = \phi_z.$$ 

This is the right direction because

$$\frac{z_2}{z_1} = \sin(\phi_z) \tan(\theta_z) \quad \quad \frac{y_2}{y_1} = \sin(\phi_y) \tan(\theta_y).$$

As a result,

$$\theta_z < \theta_y \implies \sin(\phi_z) \tan(\theta_z) < \sin(\phi_y) \tan(\theta_y) \implies \phi'_z < \phi'_y.$$ 

In our transformed coordinate system, this gives us $(0, \sin \phi_z, \cos \phi_z)$. If we transform it back to our original coordinate system, the final value is $(\sin \phi_z, \cos \phi_z, 0)$.

On the other hand, if $\theta_z > \theta_y, \phi_0 = \frac{\pi}{2}$ because $\theta_z, \theta_y \in [0, \frac{\pi}{2}]$. If we repeat the calculation from before for $a$ with $\phi_0 = \frac{\pi}{2}$, we get that $a = \cot \phi_z$. This time, $\phi' = 0$ is the right direction and so $\theta' = \frac{\pi}{2}$. As a result, in our transformed coordinate system, this gives us $(1, 0, 0)$. If we transform it back to our original coordinate system, the final value is $(0, 0, 1)$.

Then, for $\phi_z = \frac{\pi}{2}$, we can first rotate our coordinate system $\frac{\pi}{4}$ radians around the z-axis. Then, in this new coordinate system, $\phi''_z = \frac{\pi}{4}$. We can then apply the argument above and then apply the appropriate transformations to the results to get back to our original Cartesian coordinates. A similar argument holds for $\phi_z = 0$ except that we rotate it $-\frac{\pi}{4}$ radians. ✷

A.3.2 Generating the model in $\Delta^2$

The previous subsection describes how we can generate the geodesic and its endpoints if we have two points. However, we might also want to generate data according to the model. This subsection will describe how to do so.
Given a point \( p \) in the positive orthant, we can do so in the following manner. First, we can easily generate a point in a random direction \( \phi \) from \((0, 0, 1)\) because if we use spherical coordinates, we can set \( \theta \) to be \( c \) and draw \( \phi \sim \text{Unif}(0, 2\pi) \). Call this new point \( q \) in the positive orthant. After doing so, we can define a linear transform to change our axes from the standard one to one such that \((0, 0, 1)\) gets mapped the the point in the positive orthant and the other axes get mapped to vectors perpendicular to that point. Call this new coordinate system centered around \( p \), \( S^3_p \). If we apply this transformation to \( q \), we get a point in a random direction from \((0, 0, 1)\).

One way to define this linear transformation is the following. If \( r = 1, \phi_p, \) and \( \theta_p \) are the spherical coordinates for \( p \) in the positive orthant, the linear transformation from \( S^3_p \to S^3 \) can be defined by the following matrix \( O_p \):

\[
\begin{pmatrix}
\cos \theta_p \cos \phi_p & -\sin \phi_p & \sin \theta_p \cos \phi_p \\
\cos \theta_p \sin \phi_p & \cos \phi_p & \sin \theta_p \sin \phi_p \\
-\sin \theta_p & 0 & \cos \theta_p
\end{pmatrix}.
\]

(A.2)

In essence, we’re sending \((0, 0, 1)\) to \((\sin \theta_p \cos \phi_p, \sin \theta_p \sin \phi_p, \cos \theta_p)\), i.e. our point \( p, (0, 1, 0) \) to \((\sin(\theta_p + \frac{\pi}{2}) \cos \phi_p, \sin(\theta_p + \frac{\pi}{2}) \sin \phi_p, \cos(\theta_p + \frac{\pi}{2}))\), i.e. a point perpendicular to \( p \), and \((1, 0, 0)\) to \((-\sin \phi_p, \cos \phi_p, 0)\), the cross product of the first two vectors.

As a result of the procedure, we can say that the geodesic is generated by \( \phi \). After all, as the lemma below shows, the \( \phi \) for any point on the geodesic between \((0, 0, 1)\) and \( O_p^{-1} q \) must be the same \( \phi \) as \( O_p^{-1} q \)’s \( \phi \) in spherical coordinates. Because we’re considering the L2 distance on the surface of the sphere and the linear transformation can be represented by an orthogonal matrix, the linear transformation preserves distances and is a bijection. As a result, it is a global isometry and thus preserves geodesics. Thus, we can map the geodesic between \((0, 0, 1)\) and \( O_p^{-1} q \) to the geodesic on the orthant.

**Lemma A.3.3.** Suppose \( q' \) can be represented by \((\theta'_q, \phi'_q)\) in spherical coordinates. Then, the \( \phi' \) for the spherical coordinates for any point \( p' \) that lie on the geodesic between \((0, 0, 1)\) and \( q' \) must be \( \phi'_q \).

**Proof.** Our geodesic equation is not defined for \((0, 0, 1)\) because the choice of \( \phi' \) is ill-defined. As a result, we transform our two points by swapping the x and z coordinates, i.e.

\[
\begin{pmatrix}
0 \\
0 \\
1
\end{pmatrix} \mapsto \begin{pmatrix}
1 \\
0 \\
0
\end{pmatrix} \quad \begin{pmatrix}
\sin(\theta'_q) \cos(\phi'_q) \\
\sin(\theta'_q) \sin(\phi'_q) \\
\cos \theta'_q
\end{pmatrix} \mapsto \begin{pmatrix}
\cos \theta'_q \\
\sin(\theta'_q) \sin(\phi'_q) \\
\sin(\theta'_q) \cos(\phi'_q)
\end{pmatrix}.
\]
Because this transformation can be represented by the orthogonal matrix,

$$\begin{pmatrix}
0 & 0 & 1 \\
0 & 1 & 0 \\
1 & 0 & 0
\end{pmatrix},$$

and we are considering the L2 distance, the transformation is an isometry and preserves geodesics. Denote $q'$ in this new system as $q''$, $o'$ in this new system as $o''$, and any other related variables in a similar manner. Then, if we apply our geodesic equation to this new system, we get that

$$\phi''_0 = \tan^{-1}\left(\frac{0 - \sin(\theta'_q) \cos(\phi'_q)}{0 - 0}\right) = \pm \frac{\pi}{2},$$

provided $\theta'_q \neq 0, \pi$ and $\phi'_q \neq \frac{\pi}{2}, \frac{3\pi}{2}$. Provided these restrictions hold, we also get that

$$a'' = \frac{\sin(\theta'_q) \cos(\phi'_q)}{\cos(\theta'_q) \cos(\frac{\pi}{2}) + \sin(\theta'_q) \sin(\phi'_q) \sin(\frac{\pi}{2})}$$

$$= \pm \frac{\sin(\theta'_q) \cos(\phi'_q)}{\pm \sin(\theta'_q) \sin(\phi'_q)}$$

$$= \pm \cot(\phi'_q).$$

Note that if $\phi''_0 = \frac{\pi}{2}$, $a'' = \cot(\phi'_q)$ and similarly, if $\phi''_0 = -\frac{\pi}{2}$, $a'' = -\cot(\phi'_q)$. Then, under the same restrictions,

$$\cot(\theta'') = a \cos(\phi'' - \phi''_0)$$

$$\cot(\theta') = \pm \cot(\phi'_q) \cos(\phi'' - (\pm \frac{\pi}{2}))$$

$$\cot(\theta'') = \pm \cot(\phi'_q) \sin(\pm \phi'')$$

$$\phi''_3 = \cot(\phi'_q) \phi''_2.$$

If we undo our transformation, then $\phi'$ for $o'$ must be $\arctan\left(\frac{\phi''_0}{\cot(\phi'_q) \phi''_2}\right)$ or $\phi'_q$. For $\theta'_q = 0, \pi$, the geodesic is not unique. So, for consistency, we can pick the geodesic with $\phi' = \phi'_q$ in spherical coordinates. For $\phi'_q \neq \frac{\pi}{2}, \frac{3\pi}{2}$, we can again transform our data by rotating it another $\frac{\pi}{4}$ around the the z-axis in our new coordinate system. Because the rotation uses an orthogonal matrix, the geodesic is preserved and we can reapply the argument above.

Of course, to be more precise, we should say that the geodesic is generated by $\phi$ if $S^3_p \mapsto S^3$ is represented by the matrix mentioned earlier. There can be other choices to represent the
transformation and the inverse might give us a different \( \phi \) for the same point.

There is one more thing that we have to be careful about when we generate the data using this approach. In our generative model, we assign probability to the proportion of the max simplex Hellinger distance in a direction of \( \phi \). We do so to further discourage the model from choosing \( \eta \) such that points that are close to the boundary are sent to the boundary by the affine transformation. After all, a point that is close to the boundary and sent to the boundary has a distance proportion of 1. While it might be easier to generate data with respect to \( t \), the following lemma shows that there is a difference.

**Lemma A.3.4.** Suppose \( \theta_1 \) corresponds to the point on the boundary for \( g(t) \) in the direction of \( \phi \). Then, for \( t \in [0, 1] \), the proportion is equal to

\[
\frac{\sin(\theta_1 t)}{\sin(\frac{\theta_1 t}{2})}
\]

**Proof.** \( O_p \) is an orthogonal matrix and thus an isometry. Hence, the simplex Hellinger distance in three dimensions between \( p \) and \( g(t) \) in the direction of \( \phi \) is equal to

\[
\frac{1}{\sqrt{2}} \left\| \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} \sin(\theta_1 t) \cos(\phi) \\ \sin(\theta_1 t) \sin(\phi) \\ \cos(\theta_1 t) \end{pmatrix} \right\|_2
\]

Then,

\[
\frac{1}{\sqrt{2}} \left\| \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} \sin(\theta_1 t) \cos(\phi) \\ \sin(\theta_1 t) \sin(\phi) \\ \cos(\theta_1 t) \end{pmatrix} \right\|_2^2
\]

\[
= \frac{1}{\sqrt{2}} \sqrt{(\sin(\theta_1 t) \cos(\phi))^2 + (\sin(\theta_1 t) \sin(\phi))^2 + (1 - \cos(\theta_1 t))^2}
\]

\[
= \frac{1}{\sqrt{2}} \sqrt{\sin^2(\theta_1 t) \cos^2(\phi) + \sin^2(\theta_1 t) \sin^2(\phi) + (1 - 2 \cos(\theta_1 t) + \cos^2(\theta_1 t))}
\]

\[
= \frac{1}{\sqrt{2}} \sqrt{\sin^2(\theta_1 t) + 1 - 2 \cos(\theta_1 t) + \cos^2(\theta_1 t)}
\]

\[
= \sqrt{2} - 2 \cos(\theta_1 t)
\]

Because the maximum distance occurs when \( t = 1 \), the max Hellinger distance on the simplex is \( \sqrt{2} \sin(\frac{\theta_1 t}{2}) \). As a result, the proportion is \( \frac{\sin(\theta_1 t)}{\sin(\frac{\theta_1 t}{2})} \). \( \Box \)

As we see in the plots in Figure A.1, the difference doesn’t look to be significant even if it is slightly more noticeable for \( \frac{\pi}{3} \). Still, we have to be careful whether we’re generating according to
the proportion for the simplex Hellinger distance or $t$.

### A.3.3 Geodesics in higher dimensions

In order to derive the geodesic in higher dimensions, we first have to define the transformation from Cartesian coordinates to spherical coordinates in higher dimensions. In a slight modification of the notation used by Blumenson[13], if $x \in \Delta^D$, define the transformation as

$$
\begin{align*}
\sqrt{x_1} &= \cos \phi_1, \\
\sqrt{x_j} &= \cos \phi_j \prod_{k=1}^{j-1} \sin \phi_k, \quad j = 2, 3, \ldots, D, \\
\sqrt{x_{D+1}} &= \prod_{k=1}^{D} \sin \phi_k.
\end{align*}
$$

(A.3)

with $\phi_1, \phi_2, \ldots, \phi_D \in [0, \frac{\pi}{2}]$. This is a slight inversion from the 3D spherical coordinates because $\phi_1 = \theta$ and $\phi_2 = \phi$.

Then, the arc-length in these coordinates is the following.

**Lemma A.3.5.** If $\phi_1, \phi_2, \ldots, \phi_D$ represent the higher dimensional spherical coordinates, the arc-length is

$$
\int_{\phi_1} \left[ 1 + \sum_{j=2}^{d-1} \left( \prod_{k=1}^{j-1} \sin^2 \phi_k \right) \left( \frac{d\phi_j}{d\phi_1} \right)^2 \right] d\phi_1
$$

Proof. To calculate $\sum_{i=1}^{d} \left( \frac{dx_i}{dt} \right)^2$, we need to take the derivatives with respect to our definition of
the spherical coordinates given in Equation A.3.

\[
\frac{dx_1}{dt} = -\sin \phi_1 \frac{d\phi_1}{dt}
\]

\[
\frac{dx_j}{dt} = -\frac{d\phi_j}{dt}\prod_{k=1}^{j-1} \sin \phi_k + \sum_{k=1}^{j-1} \cos \phi_j \cos \phi_k \frac{d\phi_k}{dt}\prod_{l=1, l\neq k}^{j-1} \sin \phi_l \quad j = 2, 3, \ldots, d - 1
\]

\[
\frac{dx_d}{dt} = \sum_{k=1}^{d-1} \cos \phi_k \frac{d\phi_k}{dt}\prod_{l=1, l\neq k}^{d-1} \sin \phi_l.
\]

Let us first look at the \((\frac{d\phi_i}{dt})^2\) terms.

- For \(i = 1\), if we gather all the terms with \((\frac{d\phi_1}{dt})^2\), we have that

\[
\sin^2 \phi_1 + \sum_{j=2}^{d-1} \cos^2 \phi_j \cos^2 \phi_1 \prod_{l=2}^{j-1} \sin^2 \phi_l + \cos^2 \phi_1 \prod_{l=2}^{d-1} \sin^2 \phi_l
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1 \left( \sum_{j=2}^{d-1} \cos^2 \phi_j \prod_{l=2}^{j-1} \sin^2 \phi_l + \prod_{l=2}^{d-1} \sin^2 \phi_l \right)
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1 \left( \sum_{j=2}^{d-2} \cos^2 \phi_j \prod_{l=2}^{j-1} \sin^2 \phi_l + \cos^2 \phi_{d-1} \prod_{l=2}^{d-2} \sin^2 \phi_l + \prod_{l=2}^{d-1} \sin^2 \phi_l \right)
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1 \left( \sum_{j=2}^{d-2} \cos^2 \phi_j \prod_{l=2}^{j-1} \sin^2 \phi_l + \prod_{l=2}^{d-2} \sin^2 \phi_l \right) \left( \cos^2 \phi_{d-1} + \sin^2 \phi_{d-1} \right)
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1 \left( \sum_{j=2}^{d-2} \cos^2 \phi_j \prod_{l=2}^{j-1} \sin^2 \phi_l + \prod_{l=2}^{d-2} \sin^2 \phi_l \right)
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1 \left( \cos^2 \phi_2 + \sin^2 \phi_2 \right)
\]

\[
= \sin^2 \phi_1 + \cos^2 \phi_1
\]

\[
= 1.
\]

- For \(i = d - 1\), we have that

\[
\cos^2 \phi_d \prod_{l=1}^{d-2} \sin^2 \phi_l + \prod_{l=1}^{d-1} \sin^2 \phi_k = \left( \prod_{l=1}^{d-2} \sin^2 \phi_l \right) \left( \cos^2 \phi_d + \sin^2 \phi_d \right) = \prod_{l=1}^{d-2} \sin^2 \phi_l.
\]
• For \( i \neq 1, d - 1 \), if we repeat this process, we have that

\[
\prod_{l=1}^{i} \sin^2 \phi_l + \sum_{j=i+1}^{d-1} \cos^2 \phi_j \cos^2 \phi_i \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l + \cos^2 \phi_i \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l
\]

\[
= \prod_{l=1}^{i-1} \sin^2 \phi_l \left( \sin^2 \phi_i + \sum_{j=i+1}^{d-1} \cos^2 \phi_j \cos^2 \phi_i \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l + \cos^2 \phi_i \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l \right)
\]

\[
= \prod_{l=1}^{i-1} \sin^2 \phi_l \left( \sin^2 \phi_i + \cos^2 \phi_i \left( \sum_{j=i+1}^{d-1} \cos^2 \phi_j \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l + \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l \right) \right)
\]

\[
= \prod_{l=1}^{i-1} \sin^2 \phi_l \left( \sin^2 \phi_i + \cos^2 \phi_i \left( \sum_{j=i+1}^{d-1} \cos^2 \phi_i + \prod_{l=1, l \neq i}^{j-1} \sin^2 \phi_l \right) \right)
\]

\[
= \prod_{l=1}^{i-1} \sin^2 \phi_l.
\]

Now let us first look at the cross-terms, i.e. \( \left( \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \right), k_1 \neq k_2. \)

\[
\sum_{j=2}^{d-1} \left( -\frac{d\phi_j}{dt} \prod_{k=1}^{j} \sin \phi_k + \sum_{k=1}^{j-1} \cos \phi_j \cos \phi_k \frac{d\phi_k}{dt} \prod_{l=1, l \neq k}^{j-1} \sin \phi_l \right)^2 + \left( \sum_{k=1}^{d-1} \cos \phi_k \frac{d\phi_k}{dt} \prod_{l=1, l \neq k}^{d-1} \sin \phi_l \right)^2
\]

\[
= \sum_{j=2}^{d-1} 2 \left( -\frac{d\phi_j}{dt} \prod_{k=1}^{j} \sin \phi_k \right) \left( \sum_{k=1}^{j-1} \cos \phi_j \cos \phi_k \frac{d\phi_k}{dt} \prod_{l=1, l \neq k}^{j-1} \sin \phi_l \right) +
\]

\[
\sum_{j=3}^{d-1} \sum_{k_1=2}^{j-1} \sum_{k_2=1}^{k_1-1} 2 \cos^2 \phi_j \cos \phi_{k_1} \sin \phi_{k_1} \cos \phi_{k_2} \sin \phi_{k_2} \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1,k_2}^{j-1} \sin^2 \phi_l +
\]

\[
\sum_{k_1=2}^{d-1} \sum_{k_2=1}^{d-1} 2 \cos \phi_{k_1} \sin \phi_{k_1} \cos \phi_{k_2} \sin \phi_{k_2} \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1,k_2}^{d-1} \sin^2 \phi_l +
\]

\[
= \sum_{j=2}^{d-1} \sum_{k=1}^{j-1} -2 \cos \phi_j \sin \phi_j \cos \phi_k \sin \phi_k \frac{d\phi_j}{dt} \frac{d\phi_k}{dt} \prod_{l=1, l \neq k}^{j-1} \sin^2 \phi_l +
\]

128
We can simplify this using cases.

- Let us look at $k_1 = d - 1$. Then, if we gather terms, we get that

\[
\begin{align*}
&\sum_{k_2=1}^{d-2} \cos \phi_{k_2} \cos \phi_{k_2} \sin \phi_{k_2} \sin \phi_{k_2} \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-2} \sin^2 \phi_l + \\
&\sum_{k_2=1}^{d-2} \cos \phi_{k_2} \cos \phi_{k_2} \sin \phi_{k_2} \sin \phi_{k_2} \frac{d\phi_{d-1}}{dt} \frac{d\phi_{d-1}}{dt} \prod_{l=1, l \neq k_2}^{d-2} \sin^2 \phi_l + \\
&\sum_{j=k_1+1}^{d-1} \cos \phi_j \cos \phi_j \sin \phi_j \sin \phi_j \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l + \\
&\sum_{j=k_1+1}^{d-1} \cos \phi_j \cos \phi_j \sin \phi_j \sin \phi_j \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l + \\
&\sum_{j=k_1+1}^{d-1} \cos \phi_j \cos \phi_j \sin \phi_j \sin \phi_j \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l \\
&= 0.
\end{align*}
\]

- Let us look at $k_1 \neq d - 1$. 

\[
\begin{align*}
&- \sum_{j=k_1+1}^{d-1} \cos^2 \phi_j \cos \phi_j \sin \phi_j \sin \phi_j \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{k_1-1} \sin^2 \phi_l + \\
&\sum_{j=k_1+1}^{d-1} \cos \phi_j \cos \phi_j \sin \phi_j \sin \phi_j \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l + \\
&2 \cos \phi_{k_1} \cos \phi_{k_1} \sin \phi_{k_1} \sin \phi_{k_2} \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l \\
&= 2 \cos \phi_{k_1} \cos \phi_{k_1} \sin \phi_{k_1} \sin \phi_{k_2} \frac{d\phi_{k_1}}{dt} \frac{d\phi_{k_2}}{dt} \prod_{l=1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l \\
&\left( - \sin^2 \phi_{k_1-1} + \sum_{j=k_1+1}^{d-1} \cos^2 \phi_j \prod_{l=k_1-1, l \neq k_1, k_2}^{j-1} \sin^2 \phi_l + \prod_{l=k_1-1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l \right).
\end{align*}
\]
As before,

\[-\sin^2 \phi_{k_1 -1} + \sum_{j=k_1+1}^{d-1} \cos^2 \phi_j \prod_{l=k_1-1, l \neq k_1, k_2}^{j-1} \sin^2 \phi_l + \prod_{l=k_1-1, l \neq k_1, k_2}^{d-1} \sin^2 \phi_l\]

\[= -\sin^2 \phi_{k_1 -1} + \sum_{j=k_1+1}^{d-2} \cos^2 \phi_j \prod_{l=k_1-1, l \neq k_1, k_2}^{j-1} \sin^2 \phi_l + \prod_{l=k_1-1, l \neq k_1, k_2}^{d-2} \sin^2 \phi_l\]

\[= -\sin^2 \phi_{k_1 -1} + \cos^2 \phi_{k_1+1} \sin^2 \phi_{k_1 -1} + \sin^2 \phi_{k_1+1} \sin^2 \phi_{k_1 -1}\]

\[= -\sin^2 \phi_{k_1 -1} + \sin^2 \phi_{k_1 -1}\]

\[= 0.\]

Hence, the cross-terms disappear. As a result, with some rearrangement, we get the arc-length formula stated in the lemma.

From this lemma, we see that the Euler-Lagrange equation is fairly straightforward to solve in three dimensions because it only involves one function. However, in higher dimensions, the Euler-Lagrange is hard to solve because of the dependencies on previous $\phi$. Indeed, the momentum, Hamiltonian, and canonical equations, stated below without proof, are difficult to solve:

\[p_i = \sqrt{\frac{1}{1 + \sum_{j=2}^{d-1} \left( \prod_{k=1}^{j-1} \sin^2 \phi_k \right) \left( \frac{d \phi_j}{d \phi_1} \right)^2}} \prod_{k=1}^{i} \sin^2 \phi_k \frac{d \phi_{i+1}}{d \phi_1},\]

\[H = -\sqrt{1 - \sum_{i=1}^{d-2} \frac{p_i^2}{\prod_{k=1}^{i} \sin^2 \phi_k}}\]

\[\frac{dp_i}{d \phi_1} = -\frac{p_i}{H \prod_{k=1}^{i} \sin^2 \phi_k},\]

\[\frac{d \phi_{i+1}}{d \phi_1} = -p_i \left( \sum_{j=i+1}^{d-2} \frac{p_j^2}{\prod_{k=1}^{j} \sin^2 \phi_k} \frac{\cos \phi_j}{\sin^3 \phi_{j+1}} \right) \quad i \neq d - 2,\]
\[
\frac{dp_{d-2}}{d\phi_1} = 0.
\]

Luckily, we only use the geodesic equation to determine the boundary endpoint when performing inference in our model. In particular, we can use the same approach that we used to generate the geodesic in 3D. As the following lemma shows, it is straightforward to define the geodesic from \(e_1 \in \Delta^d\) to any other point.

**Lemma A.3.6.** The simplex Hellinger geodesic from \(e_1 \in \Delta^D\) to any \(y \in \Delta^D\) has \(\phi_2, \phi_3, \ldots, \phi_d\) constant.

**Proof.** For \(e_1 \in \Delta^D\) for \(i \in 2, 3, \ldots, D + 1, \phi_1 = 0\) and so we are free to set any value for \(\phi_2, \phi_3, \ldots, \phi_{d-1}\). If \(\phi_y^2, \phi_y^3, \ldots, \phi_y^{d-1}\) are the higher dimensional spherical coordinates for \(y\), set \(\phi_2 = \phi_y^2, \phi_3 = \phi_y^3, \ldots, \) and \(\phi_{d-1} = \phi_y^{d-1}\).

If we then take the path with \(\phi_2, \phi_3, \ldots, \phi_d\) constant and only vary \(\phi_1, \frac{d\phi_2}{d\phi_1} = \frac{d\phi_3}{d\phi_1} = \cdots = \frac{d\phi_{d-1}}{d\phi_1} = 0\). Then, by Lemma A.3.5, the arc-length integral becomes \(\int_{\phi_1} 1\). This clearly minimizes the integral. \(\square\)

From this lemma, we see that only \(\phi_1\) varies for that geodesic in the higher dimensional spherical coordinates. As a result, if we can use an orthogonal matrix to transform \(\sqrt{z}\) to \(e_1\) and \(\sqrt{y}\) to a point defined in that space, all we have to do is find the \(\phi_1\) for the vector in that space such that when mapped back, it is on the boundary.

This orthogonal matrix is not difficult to create. Let \(p\) be the point in the positive orthant and \(\phi_1^p, \phi_2^p, \ldots, \phi_{d-1}^p\) represent the higher dimensional spherical coordinates for \(p\). Define the orthogonal matrix \(O_p\) as following. Set the first column of \(O_p\) to be \(p\). For column \(j\), set \(\phi_k = \frac{\pi}{2}\) for \(k < j - 1, \phi_{j-1}^p + \frac{\pi}{2}\) for \(k = j - 1, \) and \(\phi_k^p\) for \(k > j - 1\) and apply the transform from higher dimensional spherical coordinates to Cartesian coordinates. As one example of this orthogonal matrix, this is the orthogonal matrix we defined in three dimensions, but with the first and third columns swapped. Another example is the orthogonal matrix in four dimensions:

\[
\begin{pmatrix}
\cos \phi_1^p & -\sin \phi_1^p & 0 & 0 \\
\sin \phi_1^p \cos \phi_2^p & \cos \phi_1^p \cos \phi_2^p & -\sin \phi_2^p & 0 \\
\sin \phi_1^p \sin \phi_2^p \cos \phi_3^p & \cos \phi_1^p \sin \phi_2^p \cos \phi_3^p & \cos \phi_2^p \cos \phi_3^p & -\sin \phi_3^p \\
\sin \phi_1^p \sin \phi_2^p \sin \phi_3^p & \cos \phi_1^p \sin \phi_2^p \sin \phi_3^p & \cos \phi_2^p \sin \phi_3^p & \cos \phi_3^p \\
\end{pmatrix}.
\]

**Lemma A.3.7.** If \(O_p\) is defined as above, it is an orthogonal matrix.

**Proof.** The last column of \(O_p\) is \(-\sin \phi_{d-1}^p\) and \(\cos \phi_{d-1}^p\) for the last two rows and zero elsewhere. Clearly, when we dot product it with itself, the dot product is 1. On the other hand, when we dot product it with any other column, the only terms left are the ones in the last two rows. By definition
of the transformation and \( O_p \), the last two rows for any other column can be written as \( c \cos \phi^p_{d-1} \) and \( c \sin \phi^p_{d-1} \) respectively. Hence, the dot product with any other column is zero.

Then, the second to last column of \( O_p \) is \(- \sin \phi^p_{d-2}, \cos \phi^p_{d-2} \cos \phi^p_{d-1}, \) and \( \cos \phi^p_{d-2} \sin \phi^p_{d-1} \) and 0 otherwise. If we dot product it with itself, we get

\[
\sin^2 \phi^p_{d-2} + \cos^2 \phi^p_{d-2} \cos^2 \phi^p_{d-1} + \cos^2 \phi^p_{d-2} \sin^2 \phi^p_{d-1} = \sin^2 \phi^p_{d-2} + \cos^2 \phi^p_{d-2} = 1.
\]

On the other hand, if we dot product it with any previous column vectors, the only terms left are the ones in the last three rows. By definition of the transformation and \( O_p \), the last three rows for any previous column can be written as \( c \cos \phi^p_{d-2}, c \sin \phi^p_{d-2} \cos \phi^p_{d-1}, \) and \( c \sin \phi^p_{d-2} \sin \phi^p_{d-1} \) respectively. Then, the dot product is

\[
- \sin \phi^p_{d-2} (c \cos \phi^p_{d-2}) + \cos \phi^p_{d-2} \cos \phi^p_{d-1} (c \sin \phi^p_{d-2} \cos \phi^p_{d-1}) + c \sin \phi^p_{d-2} \sin \phi^p_{d-1} (c \sin \phi^p_{d-2} \sin \phi^p_{d-1})
= - c \sin \phi^p_{d-2} \cos \phi^p_{d-2} + c \sin \phi^p_{d-2} \cos \phi^p_{d-2} \cos^2 \phi^p_{d-1} + c \sin \phi^p_{d-2} \cos \phi^p_{d-2} \sin^2 \phi^p_{d-1}
= - c \sin \phi^p_{d-2} \cos \phi^p_{d-2} + c \sin \phi^p_{d-2} \cos \phi^p_{d-2} \cos^2 \phi^p_{d-1}
=0.
\]

We can then repeat this argument for the third to last column, fourth to last, and so on until we have done so for the second column.

Now that we have these lemmas, we can be more precise with our notation. As before, let \( q \) be a point in the positive orthant that we are interested in extending the geodesic from \( p \) to. We hope to extend the geodesic to the point on the boundary, \( q' \). Let \( O_p \) be an orthogonal matrix, \( \mathbb{S}^d_p \) represents the space defined by \( O_p \), and \( \tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_{d-1} \) be the higher spherical dimensional coordinates in \( \mathbb{S}^d_p \). Further, let \( \tilde{\phi}_1^q, \tilde{\phi}_2^q, \ldots, \tilde{\phi}_{d-1}^q \) denote the higher dimensional spherical coordinates for \( O_p^{-1} q \) and \( \tilde{\phi}_1^q, \tilde{\phi}_2^q, \ldots, \tilde{\phi}_{d-1}^q \) denote the higher dimensional spherical coordinates for \( O_p^{-1} q' \). By doing so, we can calculate \( \tilde{\phi}_1^q \) according to the following lemma.

**Lemma A.3.8.** Suppose \( p, q \) are points in the orthant. Let \( c_1, c_2, \ldots, c_D \) be the Cartesian coordinates after applying the transformation defined in Equation A.3 to \( \tilde{\phi}_1^q, \tilde{\phi}_2^q, \ldots, \tilde{\phi}_D^q \). Then, if \( \min^+ \) denotes the smallest positive value in a set, let

\[
\tilde{\phi}_1^q = \min_k^+ \left( \pi + \tan^{-1} \left( \frac{- (O_p)_{k,1}}{\sum_{i=1}^{d-1} (O_p)_{k,i+1} c_i} \right) \mod \pi \right).
\]

**Proof.** If \( O_p \tilde{q}' = q' \), \( q' \) the point on the boundary from extending \( p \) to \( q \) and \( \tilde{q}' \in \mathbb{S}^d_p \), then \( \tilde{\phi}_2^q = \tilde{\phi}_2^q, \tilde{\phi}_3^q = \tilde{\phi}_3^q, \tilde{\phi}_{d-1}^q = \tilde{\phi}_{d-1}^q \) from Lemma A.3.6. Then, based on Equation A.3, \( \tilde{q}_1^q = \cos \phi_1 \),
\( q_2' = \sin \phi_1 c_1, \quad q_3' = \sin \phi_1 c_2, \ldots, \quad q_{D+1}' = \sin \phi_1 c_D. \) Hence,

\[
(O_p q')_k = 0
\]

\[
(O_p)_{k,1} \cos \tilde{\phi}_1' + \left( \sum_{i=1}^{D} (O_p)_{k,i+1} c_i \right) \sin \tilde{\phi}_1' = 0
\]

\[
\left( \sum_{i=1}^{D} (O_p)_{k,i+1} c_i \right) \sin \tilde{\phi}_1' = -(O_p)_{k,1} \cos \tilde{\phi}_1'
\]

\[
\tan \tilde{\phi}_1' = \frac{-(O_p)_{k,1}}{\sum_{i=1}^{D} (O_p)_{k,i+1} c_i}
\]

\[
\tilde{\phi}_1' = \tan^{-1} \left( \frac{-(O_p)_{k,1}}{\sum_{i=1}^{D} (O_p)_{k,i+1} c_i} \right).
\]

If both are 0, then any \( \tilde{\phi}_1' \) satisfies that equation so we can set \( \tilde{\phi}_1' = 0 \) in that case.

To ensure that we are in the proper range for \( \tilde{\phi}_1' \), we need to add \( \pi \) to \( \tilde{\phi}_1' \) and then mod the result by \( \pi \). Finally, because we are interested in the first \( \tilde{\phi}_1' \) going forward along the geodesic that puts \( q' \) on the boundary, we need to take the smallest positive \( \tilde{\phi}_1' \).

As a result, even though we can’t solve the geodesic equation, we can still solve for \( \tilde{\phi}_1' \) that gets us \( q' \).
APPENDIX B

Simplex Random Direction Appendix

B.1 Additional models

B.1.1 Independent random direction models

The most elementary model based on the von Mises distribution is the independent von Mises or iV model. For this model, the observations are assumed to be identically and independently distributed according to one von Mises distribution with mean $m$ and concentration parameter $\rho$. A von Mises prior on the interval $[0, 2\pi)$ with mean $u \in [0, 2\pi)$ and concentration parameter $c \in \mathbb{R}^+$ is placed on $m$. As discussed earlier in Section 3.2.1, setting $c = 0$ is equivalent to putting a uniform prior on $m$. Meanwhile, the concentration parameter $\rho$ has a Gamma prior with parameters $a \in \mathbb{R}^+$ and $b \in \mathbb{R}^+$. In summary, the iV model is the following:

$$
m \sim f_V(\cdot \mid u, c), \quad \rho \sim \Gamma(\cdot \mid a, b),
$$

$$
y_{\ell} \mid m, \rho \overset{iid}{\sim} f_V(\cdot \mid m, \rho), \quad \ell = 1, 2, \ldots, N. \tag{B.1}
$$

If there is heterogeneity in the observations, an extension of the model is the Independent von Mises Mixture or iVM model.

$$
m_{k} \sim f_V(\cdot \mid u_{k}, c_{k}), \quad k = 1, 2, \ldots, K
$$

$$
\rho_{k} \sim \Gamma(\cdot \mid a_{k}, b_{k}), \quad k = 1, 2, \ldots, K
$$

$$
\zeta_{\ell} \mid \lambda_{1}, \lambda_{2}, \ldots, \lambda_{K} \overset{iid}{\sim} \text{Cat}(\cdot \mid \lambda_{1}, \lambda_{2}, \ldots, \lambda_{K}),
$$

$$
y_{\ell} \mid \zeta_{\ell} = k, m_{k}, \rho_{k} \sim f_V(\cdot \mid m_{k}, \rho_{k}), \quad \ell = 1, 2, \ldots, N. \tag{B.2}
$$

An observation can now be distributed according to one of $K$ von Mises distribution in the interval of $[0, 2\pi)$. Each distribution has its own mean $m_k$ and concentration parameter $\rho_k$. Moreover, each distributions’ parameters are given different priors.
B.1.2 Homogeneous spatial random direction model

We proceed to modeling spatial patterns of random directions. For homogeneous spatial random direction patterns, we incorporate spatial information through the components, i.e., the parameters of the von Mises distribution. In particular, we link a Gaussian process to the mean parameters because it is a straightforward way to collate nearby observations and borrow strength. This also leads to fairly interpretable draws from the Gaussian process. Due to the von Mises distribution being concentrated at its mean, results from these models can describe the preferred random direction for a given location.

We introduce a model that applies this approach. This is the spatially varying von Mises distribution model or SvM. Each observation $y_{\ell}$ has a mean parameter, $m_\ell \in [0, 2\pi)$, and a concentration parameter, $\rho_\ell \in \mathbb{R}^+$. We use the modified arctan function defined in (3.1) to element-wise transform two draws from the Gaussian process, $Z_1, Z_2 \in \mathbb{R}^N$, into the mean parameters, $m \in [0, 2\pi)^N$. The positive-valued concentration parameter, $\rho_\ell$, is transformed via exponentiation from a real-valued $\phi_\ell$, which is then endowed with a normal prior. In summary, SvM is the following.

$$
\begin{align*}
    z_1 & \sim \text{GP}(\cdot \mid \mu_1, \Sigma), \\
    z_2 & \sim \text{GP}(\cdot \mid \mu_2, \Sigma), \\
    m & = \arctan^*(z_1, z_2), \\
    \varphi & \overset{iid}{\sim} \mathcal{N}(\cdot \mid \nu, \varsigma^2), \\
    \rho & = \exp\{\varphi\}, \\
    y_{\ell} \mid m_\ell, \rho_\ell & \sim f_V(\cdot \mid m_\ell, \rho_\ell), \quad \ell = 1, 2, \ldots, N.
\end{align*}
$$

B.2 Model properties

B.2.1 SvM model

We begin by discussing how we derive Lemma B.2.2. After using the law of total expectation and trigonometric identities, we see that calculating the properties of $Y_{\ell}$’s are reduced to that of the corresponding $m_\ell$. Because $(z_{1,\ell}, z_{2,\ell}) \sim \mathcal{N}((\mu_1, \mu_2), \sigma^2 I_2)$ and $m_\ell = \arctan^*(z_{1,\ell}, z_{2,\ell})$, $m_\ell$ is distributed according to $f_N((\mu_1, \mu_2), \sigma^2 I_2)$ [45]. We can extend the techniques in [70] to demonstrate that $m_\ell$ is equivalently distributed according to $f_N(\frac{1}{\sigma^2}(\mu_1, \mu_2), I_2)$. With this insight, we can use the discussion in Section 3.2.2 to compute the circular mean and variance. Then, we have the following lemmas.

**Lemma B.2.1.** Let $Z_1, Z_2 \sim \mathcal{N}((\mu_0 \cos(\alpha_\mu), \mu_0 \sin(\alpha_\mu)), \sigma^2 I_2)$. Let $(r, m)$ be random variables, $r \in (0, \infty)$, $m \in [0, 2\pi)$, such that $Z_1 = r \cos(m)$ and $Z_2 = r \sin(m)$. In addition, set $a =$
\( \mu_0 \cos(m - \alpha) \). Then,

\[
P_{N_2}(m \mid (\mu_0 \cos(\alpha), \mu_0 \sin(\alpha)), \sigma^2 I_2) = \phi \left( \frac{\mu_0 \sin(m - \alpha)}{\sigma} \right) \left( \phi \left( \frac{a}{\sigma} \right) + \frac{a}{\sigma} \Phi \left( \frac{a}{\sigma} \right) \right).
\]

**Proof.** Following the derivation in Appendix A of [70], we have that

\[
P_{N_2}(m \mid (\mu_0 \cos(\alpha), \mu_0 \sin(\alpha)), \sigma^2 I_2)
= \int_0^{\infty} \frac{1}{2\pi \sigma^2} \exp \left( -\frac{1}{2\sigma^2} (z_1 - \mu_0 \cos(\alpha))^2 + (z_2 - \mu_0 \cos(\alpha))^2 \right) dr
= \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{\mu_0^2}{2\sigma^2} \sin^2(\alpha) \right) \int_0^{\infty} \frac{1}{\sqrt{2\pi \sigma^2}} r \exp \left( -\frac{1}{2\sigma^2} (r - \mu_0 \cos(m - \alpha))^2 \right) dr.
\]

Applying the change of variables formula and still following the derivation in Appendix A, we have that

\[
\int_0^{\infty} \frac{1}{\sqrt{2\pi \sigma^2}} r \exp \left( -\frac{1}{2\sigma^2} (r - a)^2 \right) dr
= \int_0^{\infty} \frac{1}{\sqrt{2\pi \sigma^2}} (r - a) \exp \left( -\frac{1}{2\sigma^2} (r - a)^2 \right) dr + \int_0^{\infty} \frac{1}{\sqrt{2\pi \sigma^2}} a \exp \left( -\frac{1}{2\sigma^2} (r - a)^2 \right) dr
= \sigma \phi \left( \frac{a}{\sigma} \right) + a \Phi \left( \frac{a}{\sigma} \right).
\]

Set \( a = \mu_0 \cos(m - \alpha) \), then

\[
P_{N_2}(m \mid (\mu_0 \cos(\alpha), \mu_0 \sin(\alpha)), \sigma^2 I_2) = \phi \left( \frac{\mu_0 \sin(m - \alpha)}{\sigma} \right) \left( \phi \left( \frac{a}{\sigma} \right) + \frac{a}{\sigma} \Phi \left( \frac{a}{\sigma} \right) \right).
\]

\[ \square \]

**Lemma B.2.2.** If \( Y_\ell \) and \( Y_{\ell'} \) are generated according to SvM outlined in (B.3) with the random variables associated with them labeled accordingly and \((Z_1, Z_2) \sim N((\mu_1, \mu_2), \sigma^2 I_2)\) with \( \mu_1 = \mu_0 \cos(\alpha) \) and \( \mu_2 = \mu_0 \cos(\alpha) \), \( \mu_0 \in \mathbb{R}^+ \), \( \alpha \in [0, 2\pi) \), then

\[
\mathbb{E}(Y_\ell) = \alpha, \quad \text{(B.4)}
\]

\[
\text{Var}(Y_\ell) = 1 - \frac{I_1(\rho)}{I_0(\rho)} \left( \frac{\pi \beta}{2} \right)^2 \exp(-\beta)(I_0(\beta) + I_1(\beta)) \quad \beta = \frac{\mu_0^2}{4\sigma^2}, \quad \text{(B.5)}
\]

\[
\text{Corr}(Y_\ell, Y_{\ell'}) = \frac{\left( \frac{I_1(\rho)}{I_0(\rho)} \right)^2 \left( \mathbb{E}(\cos(m_\ell - m_{\ell'})) - \mathbb{E}(\cos(m_\ell + m_{\ell'} - 2\alpha)) \right)}{\sqrt{\left( 1 - \frac{I_2(\rho)}{I_0(\rho)} \mathbb{E}(\cos(2(m_\ell - \alpha))) \right) \left( 1 - \frac{I_2(\rho)}{I_0(\rho)} \mathbb{E}(\cos(2(m_{\ell'} - \alpha))) \right)}}. \quad \text{(B.6)}
\]
Proof. When deriving the circular mean and variance, we’ll drop the \( \ell \) subscript out of convenience. For the circular mean, i.e. the quantity in B.5, by the law of iterated expectations,

\[
E(e^{iY}) = E\left(E\left(e^{iy} \mid m\right)\right) = \frac{I_1(\rho)}{I_0(\rho)} E(e^{im}).
\]

Because

\[
PN_2(\alpha_u + \epsilon \mid (\mu_0 \cos(\alpha), \mu_0 \sin(\alpha)), \sigma^2 I_2)
= PN_2(\alpha_u - \epsilon \mid (\mu_0 \cos(\alpha), \mu_0 \sin(\alpha)), \sigma^2 I_2)
\]

by Lemma B.2.1, \( E(\sin(m - \alpha)) = 0 \). Thus, following the argument in Appendix A in [70],

\[
\arctan^* \left( \frac{E(\sin(m))}{E(\cos(m))} \right) = \arctan^* \left( \frac{\mu_0 \sin(\alpha)}{\mu_0 \cos(\alpha)} \right).
\]

As a result, the mean angle must be \( \alpha \) by definition.

For the circular variance, i.e. the quantity in B.5, we point out that if if \( m \) is the mean for \( Y \),

\[
E(\sin(Y - m)) = 0 \text{ [45].}
\]

Then, by the law of iterated expectations,

\[
E(\cos(Y - \alpha)) = E(E(\cos((Y - m) + (m - \alpha))))
= E(E(\cos(Y - m) \cos(m - \alpha) - \sin(Y - m) \sin(m - \alpha)))
= E(\cos(m - \alpha)E(\cos(Y - m)) - \sin(m - \alpha)E(\sin(Y - m)))
= \frac{I_1(\rho)}{I_0(\rho)} E(\cos(m - \alpha)).
\]

For \( \alpha = 0 \) and \( \sigma = 1 \), we have that \( E(\cos(m)) = (\frac{\pi \beta}{2})^{\frac{1}{2}} \exp(-\beta)(I_0(\beta) + I_1(\beta)) \) for \( \beta = \frac{\mu_0^2}{4} \). In other words, we have that for \( a = \mu_0 \cos(m) \), then

\[
\int_0^{2\pi} \cos(m) \phi(\mu_0 \sin(m))(\phi(a) + a \Phi(a)) = \left(\frac{\pi \beta}{2}\right)^{\frac{1}{2}} \exp(-\beta)(I_0(\beta) + I_1(\beta)).
\]

By change of variable, \( E(\cos(m - \alpha)) = (\frac{\pi \beta}{2})^{\frac{1}{2}} \exp(-\beta)(I_0(\beta) + I_1(\beta)) \) when \( \alpha \neq 0 \) and \( \sigma = 1 \). Finally, when \( \sigma \neq 1 \), we can define \( \tilde{\mu}_0 = \frac{m_0}{\sigma} \) based on Lemma B.2.1. Hence, \( E(\cos(m - \alpha)) = (\frac{\pi \beta}{2})^{\frac{1}{2}} \exp(-\tilde{\beta})(I_0(\tilde{\beta}) + I_1(\tilde{\beta})) \) with \( \tilde{\beta} = \frac{\mu_0^2}{4\sigma^2} \).
By definition, the circular correlation we are interested in calculating is
\[
\frac{E(\sin(\gamma - \alpha) \sin(\gamma' - \alpha))}{\sqrt{E(\sin^2(\gamma - \alpha)) E(\sin^2(\gamma' - \alpha))}}
\]
Thus, we need to compute \(E(\sin(\gamma - \alpha) \sin(\gamma' - \alpha))\) and without loss of generality, \(E(\sin^2(\gamma - \alpha))\). To compute these quantities, we will use trigonometric identities and the following fact. Because the von Mises distribution is unimodal and symmetric, it is straightforward to show that \(E(\sin(2(y_1 - m_\ell))) = 0\) as well.
Using the half angle identity,
\[
E(\sin^2(y_1 - \alpha)) = \frac{1}{2} (1 - E(\cos(2(y_1 - \alpha)))).
\]
Based on what we discussed earlier and the definition of the Bessel function,
\[
E(\cos(2(Y_\ell - \alpha))) = E(\mathbb{E}(\cos(2(Y_\ell - m_\ell) + 2(m_\ell - \alpha)) \mid m_\ell))
= E(\mathbb{E}(\cos(2(Y_\ell - m_\ell)) \cos(2(m_\ell - \alpha)) \mid m_\ell))
= \frac{I_2(\rho)}{I_0(\rho)} E(\cos(2(m_\ell - \alpha))).
\]
We can repeat this argument to compute \(E(\sin^2(Y_\ell' - \alpha))\).
Meanwhile, we can calculate \(E(\sin(\gamma - \alpha) \sin(\gamma' - \alpha))\) using the fact that
\[
E(\sin(\gamma - \alpha) \sin(\gamma' - \alpha)) = \frac{1}{2} E(\cos(Y_\ell - Y_\ell') - \cos(Y_\ell + Y_\ell' - 2\alpha)).
\]
For the former, we can use conditional independence to get that
\[
E(\cos(Y_\ell - Y_\ell')) = E(\mathbb{E}(\cos(Y_\ell - m_\ell + m_\ell' - Y_\ell' + m_\ell - m_\ell') \mid m_\ell, m_\ell'))
= E(\mathbb{E}(\cos(Y_\ell - m_\ell) \cos(m_\ell' - Y_\ell') \cos(m_\ell - m_\ell') \mid m_\ell, m_\ell'))
= \left(\frac{I_1(\rho)}{I_0(\rho)}\right)^2 E(\cos(m_\ell - m_\ell')).
\]
A similar argument holds for \(E(\cos(Y_\ell + Y_\ell - 2\alpha))\) except that
\[
E(\cos(Y_\ell + Y_\ell - 2\alpha)) = \left(\frac{I_1(\rho)}{I_0(\rho)}\right)^2 E(\cos(m_\ell + m_\ell' - 2\alpha)).
\]
Note that it becomes challenging to explicitly compute the circular correlation. Given the pro-
jected normal distribution for \( m \), it is difficult to extend our previous calculations for \( \cos(2(m - \alpha)) \) because the density involves \( \cos(m - \alpha) \) and \( \sin(m - \alpha) \), leading to a mismatch. Further, it is tricky to derive the projected normal distribution for \( m_\ell \) and \( m_{\ell'} \) because of the correlation between \( Z_{\ell,1} \) and \( Z_{\ell',1} \), and the correlation between \( Z_{\ell,2} \) and \( Z_{\ell',2} \). This makes it hard to integrate out \( r_\ell \) and \( r_{\ell'} \) if \( Z_{\ell,1} = r_\ell \cos(m_\ell) \), \( Z_{\ell,1} = r_{\ell'} \cos(m_{\ell'}) \), \( Z_{\ell',1} = r_\ell \sin(m_\ell) \), and \( Z_{\ell',2} = r_{\ell'} \sin(m_{\ell'}) \) for \( r_\ell, r_{\ell'} \in \mathbb{R}^+ \). Fortunately, it is straightforward to compute these values by simulation. Wang and Gelfand briefly explored this in their paper on projected Gaussian processes [71].

### B.2.2 SvM-p model

**Lemma B.2.3.** If \( Y_\ell \) and \( Y_{\ell'} \) are generated according to SvM-p specified in (3.16) with the random variables associated with them labeled accordingly, \( \arctan^* \) is defined in (3.1), \( \alpha = \mathbb{E}(Y_\ell) \), which is given in (3.23), and \( s(\zeta_\ell) = 1 - \sum_k P(\zeta_\ell = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(2(m_k - \alpha)) \)

\[
\mathbb{E}(Y_\ell) = \arctan^* \left( \frac{\sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \sin(m_k)}{\sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(m_k)} \right),
\]

(B.7)

\[
\text{Var}(Y_\ell) = 1 - \sum_k P(\zeta = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(m_k - \alpha),
\]

(B.8)

\[
\text{Corr}(Y_\ell, Y_{\ell'}) = \frac{1}{\sqrt{s(\zeta_\ell)s(\zeta_{\ell'}')}} \sum_{k,k'} \sum_k P(\zeta_\ell = k, \zeta_{\ell'} = k') \frac{I_1(\rho_k)}{I_0(\rho_k)} \frac{I_1(\rho_{k'})}{I_0(\rho_{k'})} (\cos(m_k - m_{k'}) - \cos(m_k + m_{k'} - 2\alpha)).
\]

(B.9)

**Proof.** Using the law of iterated expectations, we have that

\[
\mathbb{E}(e^{iY_\ell}) = \mathbb{E}\left( \mathbb{E}(e^{iY_\ell} | \zeta_\ell) \right) = \sum_k \frac{I_1(\rho_k)}{I_0(\rho_k)} e^{im_k} P(\zeta_\ell = k)
\]

By converting \( e^{im_1} \) and \( e^{im_1} \) to their respective sin and cos formulation and use a modified arctan function to get the angle between 0 and \( 2\pi \), we get the result in B.7.
From the definition of circular variance, we have that

\[
\text{Var}(Y_\ell) = 1 - \mathbb{E}\left(\cos(Y_\ell - \alpha)\right)
\]

\[
= 1 - \mathbb{E}\left(\mathbb{E}\left(\cos(Y_\ell - \alpha) \mid \zeta_\ell\right)\right)
\]

\[
= 1 - \sum_k \mathbb{E}\left(\cos((Y_\ell - m_k) + (m_k - \alpha)) \mid \zeta_\ell = k\right) P(\zeta_\ell = k)
\]

\[
= 1 - \sum_k P(\zeta_\ell = k) \frac{I_1(\rho_k)}{I_0(\rho_k)} \cos(m_k - \alpha).
\]

Finally, for the circular correlation, we first derive \( \mathbb{E}\left(\sin^2(y_\ell - \alpha)\right) \) without loss of generality. As before, we need to compute \( \mathbb{E}\left(\cos(2(y_\ell - \alpha))\right) \). Breaking it into cases as before, we have that

\[
\mathbb{E}\left(\cos(2(y_\ell - \alpha))\right) = \sum_k P(\zeta_\ell = k) \mathbb{E}\left(\cos(2((y_\ell - m_k) + (m_k - \alpha))) \mid \zeta_\ell = k\right)
\]

\[
= \sum_k P(\zeta_\ell = k) \mathbb{E}\left(\cos(2((y_\ell - m_k))) \cos(2(m_k - \alpha)) \mid \zeta_\ell = k\right)
\]

\[
= \sum_k P(\zeta_\ell = k) \frac{I_2(\rho_k)}{I_0(\rho_k)} \cos(2(m_k - \alpha)).
\]

We can repeat this calculation for \( \mathbb{E}\left(\sin^2(y_\ell - \alpha)\right) \).

For the numerator, we have that

\[
\mathbb{E}\left(\sin(y_\ell - \alpha) \sin(y_\ell' - \alpha)\right) = \frac{1}{2}\left(\mathbb{E}\left(\cos(y_\ell - \alpha - (y_\ell' - \alpha))\right) - \cos(y_\ell - \alpha + y_\ell' - \alpha))\right)
\]

\[
= \frac{1}{2}\left(\mathbb{E}\left(\cos(y_\ell - y_\ell')\right) - \cos(y_\ell + y_\ell' - 2\alpha))\right).
\]

Let us derive each part separately. For \( \mathbb{E}\left(\cos(y_\ell - y_\ell')\right) \), we can again use the same tricks.

\[
\mathbb{E}\left(\cos(y_\ell - y_\ell')\right) = \mathbb{E}\left(\mathbb{E}\left(\cos(y_\ell - y_\ell') \mid \zeta_\ell, \zeta_\ell'\right)\right)
\]

\[
= \sum_k \sum_{k'} P(\zeta_\ell = k, \zeta_\ell' = k') \mathbb{E}\left(\cos((y_\ell - m_k) + (m_k - y_\ell')) \mid \zeta_\ell = k, \zeta_\ell' = k'\right)
\]

\[
= \sum_k \sum_{k'} P(\zeta_\ell = k, \zeta_\ell' = k') \frac{I_1(\rho_k)}{I_0(\rho_k)} \mathbb{E}\left(\cos(y_\ell' - m_{k'}) \cos(m_{k'} - m_k) \mid \zeta_\ell' = k'\right)
\]

\[
= \sum_k \sum_{k'} P(\zeta_\ell = k, \zeta_\ell' = k') \frac{I_1(\rho_k)}{I_0(\rho_k)} \frac{I_1(\rho_{k'})}{I_0(\rho_{k'})} \cos(m_{k'} - m_k).
\]
We apply the same tricks for $E(\cos(y_\ell + y_{\ell'}) - 2\alpha)$.

\[
E(\cos(y_\ell + y_{\ell'} - 2\alpha)) = E(E(\cos((y_\ell - m_k) + (y_{\ell'} + m_k - 2\alpha)) | \zeta_\ell = k, \zeta_{\ell'} = k')) = \sum_k \sum_{k'} P(\zeta_\ell = k, \zeta_{\ell'} = k') E(\cos(y_\ell - m_k) \cos(y_{\ell'} + m_k - 2\alpha) | \zeta_{\ell'} = 2) = \sum_k \sum_{k'} P(\zeta_\ell = k, \zeta_{\ell'} = k') I_1(\rho_k) \frac{I_1(\rho_{k'})}{I_0(\rho_k)} \cos(m_{k'} + m_k - 2\alpha).
\]

Combining these quantities gives us the circular correlation.

B.2.3 SvM-p-2 model

We first prove the useful lemmas mentioned in the part on SvM-p-2 from Section 3.3.2.

Lemma B.2.4. If $Z \sim N(0, 1)$, $E(\psi^{-1}(Z)) = \frac{1}{2}$.

Proof. We can also show that $E(\psi^{-1}(Z)) - E(f(Z)) = 0$. For $z > z_\epsilon$, we have that

\[
\int_{z_\epsilon}^{\infty} \psi^{-1}(z) - f(z)\phi(z)dz = \int_{z_\epsilon}^{\infty} (\psi^{-1}(z) - 1)\phi(z)dz
\]

\[
= -\int_{z_\epsilon}^{\infty} (1 - \psi^{-1}(z))\phi(z)dz
\]

\[
= -\int_{z_\epsilon}^{\infty} \psi^{-1}(-z)\phi(z)dz
\]

\[
= -\int_{-\infty}^{-z_\epsilon} \psi^{-1}(z)\phi(z)dz.
\]

Hence, this cancels out $z < z_\epsilon$. A similar argument holds for $z \in [-z_\epsilon, 0]$ and $z \in [0, z_\epsilon]$. As a result, we can compute $E(\psi^{-1}(Z))$ using $E(f(Z))$. Then, we have that

\[
E(f(Z)) = \int_{-z_\epsilon}^{z_\epsilon} \frac{1}{2z_\epsilon} (z + z_\epsilon) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) + \int_{z_\epsilon}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)
\]

\[
= \frac{1}{2} (\Phi(z_\epsilon) - \Phi(-z_\epsilon)) + 1 - \Phi(z_\epsilon)
\]

\[
= \frac{1}{2}.
\]
While the calculations for the circular mean and variance are clear, the circular correlation requires some work to derive.

**Lemma B.2.5.** The circular correlation for the SvM-p-2 model described in (3.17) is

\[
\frac{2\mathbb{E} (\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha))}{\sqrt{2\mathbb{E} (\sin^2(y_\ell - \alpha)) 2\mathbb{E} (\sin^2(y_{\ell'} - \alpha))}}
\]

where

\[
\mathbb{E} (\sin^2(y_\ell - \alpha)) = \frac{1}{2} \left( 1 - P(\zeta_\ell = 1) \frac{I_2(\rho_1)}{I_0(\rho_1)} \cos(2(m_1 - \alpha)) - P(\zeta_\ell = 2) \frac{I_2(\rho_2)}{I_0(\rho_2)} \cos(2(m_2 - \alpha)) \right),
\]

\[
\mathbb{E} (\sin^2(y_{\ell'} - \alpha)) = \frac{1}{2} \left( 1 - P(\zeta_{\ell'} = 1) \frac{I_2(\rho_1)}{I_0(\rho_1)} \cos(2(m_1 - \alpha)) - P(\zeta_{\ell'} = 2) \frac{I_2(\rho_2)}{I_0(\rho_2)} \cos(2(m_2 - \alpha)) \right),
\]

\[
\mathbb{E} (\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha)) = P(\zeta_\ell, \zeta_{\ell'} = 1) \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \right)^2 (1 - \cos(2(m_1 - \alpha))) + P(\zeta_\ell, \zeta_{\ell'} = 2) \left( \frac{I_1(\rho_2)}{I_0(\rho_2)} \right)^2 (1 - \cos(2(m_2 - \alpha))) + (P(\zeta_\ell = 1, \zeta_{\ell'} = 2) + P(\zeta_\ell = 2, \zeta_{\ell'} = 1)) \frac{I_1(\rho_1) I_1(\rho_2)}{I_0(\rho_1) I_0(\rho_2)} (\cos(m_1 - m_2) - \cos(m_1 + m_2 - 2\alpha)).
\]

**Proof.** We first derive \( \mathbb{E} (\sin^2(y_\ell - \alpha)) \) without loss of generality. Like before, we need to compute \( \mathbb{E} (\cos(2(y_\ell - \alpha))) \). Breaking it into two cases, we have that

\[
\mathbb{E} (\cos(2(y_\ell - \alpha))) = P(\zeta_\ell = 1) \mathbb{E} (\cos(2((y_\ell - m_1) + (m_1 - \alpha))) | \zeta_\ell = 1) + P(\zeta_\ell = 2) \mathbb{E} (\cos(2((y_\ell - m_2) + (m_2 - \alpha))) | \zeta_\ell = 2)
\]

\[
= P(\zeta_\ell = 1) \mathbb{E} (\cos(2((y_\ell - m_1))) \cos(2(m_1 - \alpha)) | \zeta_\ell = 1) + P(\zeta_\ell = 2) \mathbb{E} (\cos(2((y_\ell - m_2))) \cos(2(m_2 - \alpha)) | \zeta_\ell = 2)
\]

\[
= P(\zeta_\ell = 1) \frac{I_2(\rho_1)}{I_0(\rho_1)} \cos(2(m_1 - \alpha)) + P(\zeta_\ell = 2) \frac{I_2(\rho_2)}{I_0(\rho_2)} \cos(2(m_2 - \alpha)).
\]

We can repeat this calculation for \( \mathbb{E} (\sin^2(y_{\ell'} - \alpha)) \).
Note that
\[ \mathbb{E}(\sin(y - \alpha) \sin(y' - \alpha)) = \frac{1}{2}(\mathbb{E}(\cos(y - \alpha - (y' - \alpha)) - \cos(y - \alpha + y' - \alpha))). \]

Meanwhile, \( \mathbb{E}(\cos(y - y')) \) breaks down into four cases:

1. \( \zeta, \zeta' = 1 \)
2. \( \zeta, \zeta' = 2 \)
3. \( \zeta = 1, \zeta' = 2 \)
4. \( \zeta = 2, \zeta' = 1 \)

For the first case,
\[
\mathbb{E}(\cos(y_1 - y_2) | \zeta, \zeta' = 1) = \mathbb{E}(\cos((y_1 - m_1) + (m_1 - y_2)) | \zeta = 1, \zeta' = 1) \\
= \mathbb{E}(\cos(y_1 - m_1) \cos(y_2 - m_1) | \zeta = 1, \zeta' = 1) \\
= \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \right)^2.
\]

This is weighted by \( P(\zeta = 1, \zeta' = 1) \). By a similar calculation, \( \mathbb{E}(\cos(y_1 - y_2) | \zeta, \zeta' = 2) = \left( \frac{I_1(\rho_2)}{I_0(\rho_2)} \right)^2. \)

Finally, the computation for the last two cases are identical. Without loss of generality, we have that
\[
\mathbb{E}(\cos(y_1 - y_2) | \zeta = 1, \zeta' = 2) = \mathbb{E}(\cos((y_1 - m_1) + (m_1 - y_2)) | \zeta = 1, \zeta' = 2) \\
= \mathbb{E}(\cos(y_1 - m_1) \cos(y_2 - m_1) | \zeta = 1, \zeta' = 2) \\
= \frac{I_1(\rho_1)}{I_0(\rho_1)} \mathbb{E}(\cos(y_2 - m_2) \cos(m_2 - m_1) | \zeta' = 2) \\
= \frac{I_1(\rho_1) I_1(\rho_2)}{I_0(\rho_1) I_0(\rho_2)} \cos(m_2 - m_1).
\]

This is weighted by \( P(\zeta = 1, \zeta' = 2) \) and \( P(\zeta = 2, \zeta' = 1) \). If we put together these four cases using the law of iterated expectations, we get \( \mathbb{E}(\cos(y_l - y_{l'})). \)

We can also modify these calculations for \( \mathbb{E}(\cos(y_l + y_{l'} - 2\alpha)) \). In cases one and two, \( m_1 - m_1 \) and \( m_2 - m_2 \) have to be added in twice. As a result, without loss of generality, we have for the first case that
\[
\mathbb{E}(\cos(y_{l} + y_{l'} - 2\alpha) | \zeta, \zeta' = 1) = \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \right)^2 \cos(2(m_1 - \alpha)).
\]
For the last two cases, we have to insert $m_1 - m_1$ and $m_2 - m_2$. Hence,

$$
\mathbb{E} (\cos(y_\ell + y_{\ell'} - 2\alpha) \mid \zeta_\ell = 1, \zeta_{\ell'} = 2) = \frac{I_1(\rho_1)}{I_0(\rho_1)} \frac{I_1(\rho_2)}{I_0(\rho_2)} \cos(m_1 + m_2 - 2\alpha).
$$

If we then weight these quantities with their region’s probability and thus use the law of iterated expectations, we get $\mathbb{E} (\cos(y_\ell + y_{\ell'} - 2\alpha))$.

Next, for the circular correlation in (3.29), we have that

$$
2\mathbb{E} (\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha)) = 
\frac{I_1(\rho_1)}{I_0(\rho_1)} \frac{I_1(\rho_2)}{I_0(\rho_2)} (\cos(m_1 - m_2) - \cos(m_1 + m_2 - 2\alpha)) + 
\mathbb{E} (\psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'})) \left( \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \right)^2 (1 - \cos(2(m_1 - \alpha))) + 
\left( \frac{I_1(\rho_2)}{I_0(\rho_2)} \right)^2 (1 - \cos(2(m_2 - \alpha))) - 
2 \frac{I_1(\rho_1)}{I_0(\rho_1)} \frac{I_1(\rho_2)}{I_0(\rho_2)} (\cos(m_1 - m_2) - \cos(m_1 + m_2 - 2\alpha)) \right).
$$

Note that this makes what we mentioned for (3.25) more clear. If we look at the part attached to $\mathbb{E} (\psi^{-1}(z_1) \psi^{-1}(z_2))$, it is the difference between the concentration of two observations belonging to the same component and the concentration of two observations belonging to different components. The part that is not attached is the concentration of two observations belonging to two different components. Intuitively, $\mathbb{E} (\psi^{-1}(z_1) \psi^{-1}(z_2))$ will be close to 1 for "nearby" observations so the concentration of two observations belonging to the same component will dominate.

In order to be more precise, one needs to compute $\mathbb{E} (\psi^{-1}(Z_\ell) \psi^{-1}(Z_{\ell'}))$. This is intractable. Instead, we may use $\mathbb{E} (f(Z_\ell) f(Z_{\ell'}))$ to bound $\mathbb{E} (\psi^{-1}(Z_\ell) \psi^{-1}(Z_{\ell'}))$. However, in order to do so, we need the following technical lemmas.

**Lemma B.2.6.** If $Z_\ell, Z_{\ell'} \sim N(0, \Sigma)$ and $s = \text{Cov}(Z_\ell, Z_{\ell'})$ and $z_\ell \in (0, 2]$,

$$
\int_{-z_\ell}^{z_\ell} \int_{z_\ell}^{\infty} \frac{1}{\sqrt{2\pi(1 - s^2)}} \exp\left(-\frac{(z_{\ell'} - s\ell)^2}{2(1 - s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_{\ell'}^2}{2}\right) \, dz_{\ell'} \, dz_\ell = 
\phi(z_\ell) \left( \Phi \left( -z_\ell \sqrt{\frac{1 - s}{1 + s}} \right) - \Phi \left( -z_{\ell'} \sqrt{\frac{1 + s}{1 - s}} \right) \right) + 
\phi(z_{\ell'}) \left( \Phi \left( \frac{z_\ell - \vartheta}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{-z_\ell - \vartheta}{\sqrt{1 - s^2}} \right) \right),
$$

where $\vartheta = \frac{(1-s^2)z_\ell}{s}$.
Proof. Using integration of parts, we have that
\[
\int_{-z_e}^{z_e} \int_{z_e}^\infty z_\ell \frac{1}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(z_\ell - s z_\ell)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dz_\ell dz_e
\]
\[
= \int_{-z_e}^{z_e} z_\ell \left(1 - \Phi\left(\frac{z_\ell - z_e}{\sqrt{1-s^2}}\right)\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)
\]
\[
= \int_{-z_e}^{z_e} z_\ell \left(\Phi\left(\frac{sz_\ell - z_e}{\sqrt{1-s^2}}\right)\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)
\]
\[
= - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_e^2}{2}\right) \Phi\left(\frac{sz_\ell - z_e}{\sqrt{1-s^2}}\right)_{-z_e}^{z_e}
\]
\[
+ \int_{-z_e}^{z_e} \frac{s}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(sz_\ell - z_e)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)
\]
\[
= - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_e^2}{2}\right) \Phi\left(\frac{sz_\ell - z_e}{\sqrt{1-s^2}}\right)_{-z_e}^{z_e}
\]
\[
+ \int_{-z_e}^{z_e} \frac{s}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(z_\ell - z_e)^2}{2\frac{1-s^2}{s^2}}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right).
\]

For the first half, we have that
\[
- \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_e^2}{2}\right) \Phi\left(\frac{sz_\ell - z_e}{\sqrt{1-s^2}}\right)_{-z_e}^{z_e}
\]
\[
= - \phi(z_e) \phi\left(\frac{sz_\ell - z_e}{\sqrt{1-s^2}}\right) + \phi(-z_e) \phi\left(\frac{-sz_\ell - z_e}{\sqrt{1-s^2}}\right)
\]
\[
= - \phi(z_e) \left(\Phi\left(\frac{(s-1)z_e}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-(s+1)z_e}{\sqrt{1-s^2}}\right)\right)
\]
\[
= - \phi(z_e) \left(\Phi\left(\frac{(s-1)z_e}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-s+1}{\sqrt{1-s^2}}\right)\right)
\]
\[
= - \phi(z_e) \left(\Phi\left(\frac{-z_e\sqrt{(1-s)^2}}{1-s^2}\right) - \Phi\left(\frac{-z_e\sqrt{(1+s)^2}}{1-s^2}\right)\right)
\]
\[
= - \phi(z_e) \left(\Phi\left(\frac{-z_e\sqrt{1-s}}{1-s}\right) - \Phi\left(\frac{z_e\sqrt{1+s}}{1-s}\right)\right).
\]

For the second half, we have the integral of the product of two Gaussian distributions. Hence, the mean is
\[
\frac{1-s^2}{s^2} \cdot \frac{z_e}{s} + (1)(0) = \frac{(1-s^2)z_e}{s(s^2 + (1-s^2))} = \frac{(1-s^2)}{s} z_e
\]
and the standard deviation is
\[
\sqrt{\frac{1-s^2}{s^2 + (1-s^2)}} = \sqrt{\frac{1-s^2}{s^2 + (1-s^2)}} = \sqrt{1-s^2}.
\]

Meanwhile, for the scaling term, we have that the mean is essentially zero and the standard deviation is
\[
\sqrt{1 + \frac{1-s^2}{s^2}} = \sqrt{\frac{s^2 + 1-s^2}{s^2}} = \sqrt{\frac{1}{s^2}} = \frac{1}{s}.
\]
We then standardize using the mean and standard deviation to get the quantity in the lemma. □

**Lemma B.2.7.** If \( Z_\ell, Z_{\ell'} \sim \mathcal{N}(0, \Sigma) \) and \( s = \text{Cov}(Z_\ell, Z_{\ell'}) \) and \( z_\epsilon \in (0, 2] \),
\[
\int_{-z_\epsilon}^{z_\epsilon} \int_{-z_\epsilon}^{z_\epsilon} z_{\ell} z_{\ell'} \frac{1}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(z_{\ell'} - s z_\ell)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dz_{\ell'} dz_\ell =
\]
\[
\frac{1-s^2}{s} \phi(z_\epsilon) \vartheta \left( \Phi\left(\frac{z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) + \Phi\left(\frac{-z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right)\right)
\]
\[
\left( \Phi\left(\frac{z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right)\right)^2
\]
\[
s\phi(z_\epsilon) \left( -z_\epsilon \vartheta \left( \Phi\left(\frac{z_\epsilon - \vartheta}{1-s^2}\right) + \Phi\left(\frac{-z_\epsilon - \vartheta}{1-s^2}\right)\right) + \Phi\left(\frac{z_\epsilon - \vartheta}{1-s^2}\right) - \Phi\left(\frac{-z_\epsilon - \vartheta}{1-s^2}\right)\right)
\]
\[
\sqrt{1-s^2} \left( \vartheta \left( \Phi\left(\frac{z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right)\right) + \vartheta \left( \Phi\left(\frac{z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right)\right)\right)
\]
\[
P(z_\ell \in [-z_\epsilon, z_\epsilon], z_{\ell'} \leq z_\epsilon \mid 0, \Sigma),
\]
where \( \vartheta = \frac{(1-s^2) z_\epsilon}{s} \).

**Proof.** To set up change of variable, we have that
\[
\int_{-z_\epsilon}^{z_\epsilon} \int_{-z_\epsilon}^{z_\epsilon} z_{\ell} z_{\ell'} \frac{1}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(z_{\ell'} - s z_\ell)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dz_{\ell'} dz_\ell
\]
\[
= \int_{-z_\epsilon}^{z_\epsilon} \int_{-z_\epsilon}^{z_\epsilon} (z_{\ell'} - s z_\ell) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_{\ell'} - s z_\ell)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dz_{\ell'} dz_\ell + \quad (B.10)
\]
\[
\int_{-z_\epsilon}^{z_\epsilon} \int_{-z_\epsilon}^{z_\epsilon} z_{\ell}(s z_\ell) \frac{1}{\sqrt{2\pi(1-s^2)}} \exp\left(-\frac{(z_{\ell'} - s z_\ell)^2}{2(1-s^2)}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dz_{\ell'} dz_\ell.
\]
Again, let us work on the two pieces of Equation B.10. Let $\tilde{z}_\epsilon = \frac{(z_\epsilon - sz_\ell)^2}{2(1-s^2)}$ and $\tilde{z}_{-\epsilon} = \frac{(z_\epsilon - sz_\ell)^2}{2(1-s^2)}$. Then, by change of variable,

$$
\int_{-z_\epsilon}^{z_\epsilon} \int_{-z_{-\epsilon}}^{z_{-\epsilon}} z_\ell \sqrt{\frac{1-s^2}{2\pi}} \exp(-u) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) dudz_\ell

= \int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{1-s^2}{2\pi}} (\exp(\tilde{z}_\epsilon) - \exp(\tilde{z}_{-\epsilon})) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)

= \int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{1-s^2}{2\pi}} \exp(\tilde{z}_\epsilon) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) +

\int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{1-s^2}{2\pi}} \exp(\tilde{z}_{-\epsilon}) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right).

(B.11)

We can rearrange $\tilde{z}_\epsilon$ to be $\frac{(z_\epsilon - sz_\ell)^2}{2(1-s^2)}$ and $\tilde{z}_{-\epsilon}$ to be $\frac{(sz_\ell - z_\epsilon)^2}{2(1-s^2)}$. By doing so, we see that we again have the product of two Gaussians for both parts of Equation B.11.

If we look at the left part of Equation B.11, our scaling term is again $\phi(z_\epsilon)$ and that the mean and standard deviation of the Gaussian are the same as that in Lemma B.2.6. As a result, we are taking the expectation of an unnormalized truncated Gaussian. Hence,

$$
\int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{1-s^2}{2\pi}} \exp(\tilde{z}_\epsilon) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)

= \frac{1-s^2}{s} \int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{s}{2\pi(1-s^2)}} \exp(\tilde{z}_\epsilon) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right)

= -\frac{1-s^2}{s} \phi(z_\epsilon) \left( \left( \Phi\left(\frac{z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) \right) \vartheta - \sqrt{1-s^2} \left( \phi\left(\frac{z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) - \phi\left(\frac{-z_\epsilon - \vartheta}{\sqrt{1-s^2}}\right) \right) \right).

$$

Meanwhile for the right part of Equation B.11, our scaling term is $\phi(-z_\epsilon) = \phi(z_\epsilon)$. While the standard deviation is the same as before, the mean is $-\vartheta$. As a result,

$$
\int_{-z_\epsilon}^{z_\epsilon} -z_\ell \sqrt{\frac{1-s^2}{2\pi}} \exp(\tilde{z}_\epsilon) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_\ell^2}{2}\right) =

\frac{1-s^2}{s} \phi(z_\epsilon) \left( \left( \Phi\left(\frac{z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) - \Phi\left(\frac{-z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) \right) \vartheta + \sqrt{1-s^2} \left( \phi\left(\frac{z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) - \phi\left(\frac{-z_\epsilon + \vartheta}{\sqrt{1-s^2}}\right) \right) \right).
$$
Hence, combining the two, we get that for the left part of Equation B.10,

$$\frac{1 - s^2}{s} \phi(z_\varepsilon) \vartheta \left( \Phi \left( \frac{z_\varepsilon + \vartheta}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{-z_\varepsilon + \vartheta}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{z_\varepsilon - \vartheta}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{-z_\varepsilon - \vartheta}{\sqrt{1 - s^2}} \right) \right).$$

(B.12)

For the right part, we have that

$$\int_{-z_\ell}^{z_\ell} \int_{-z_\ell}^{z_\ell} z_\ell (sz_\ell) \frac{1}{\sqrt{2\pi (1 - s^2)}} \exp \left\{ \frac{-(s_\ell^2 - sz_\ell^2)}{2(1 - s^2)} \right\} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) dz_\ell dz_\ell$$

$$= \int_{-z_\ell}^{z_\ell} s_\ell^2 \left( \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{-z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) dz_\ell$$

$$= \int_{-z_\ell}^{z_\ell} s_\ell^2 \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) - \int_{-z_\ell}^{z_\ell} \Phi \left( \frac{-z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) dz_\ell.$$  

(B.13)

Let us focus on the right part of Equation B.13. By integration of parts, we have that

$$\int_{-z_\ell}^{z_\ell} s_\ell^2 \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right)$$

$$= - s \left( z_\ell \frac{-s}{\sqrt{2\pi (1 - s^2)}} \exp \left( \frac{-(sz_\ell - z_\varepsilon^2)}{2(1 - s^2)} \right) + \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) \bigg|_{-z_\ell}^{z_\ell} +$$

$$\int_{-z_\ell}^{z_\ell} s \left( z_\ell \frac{-s}{\sqrt{2\pi (1 - s^2)}} \exp \left( \frac{(sz_\ell - z_\varepsilon^2)}{2(1 - s^2)} \right) + \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right)$$

$$= - s \phi(z_\varepsilon) \left( -z_\ell \vartheta \left( \frac{1 - s}{1 + s} \right) + \Phi \left( \frac{z_\ell \sqrt{1 - s}}{1 + s} \right) \right) +$$

$$s \phi(z_\varepsilon) \left( z_\ell \vartheta \left( \frac{1 + s}{1 - s} \right) + \Phi \left( \frac{z_\ell \sqrt{1 + s}}{1 - s} \right) \right) +$$

$$\int_{-z_\ell}^{z_\ell} sz_\ell \frac{-s}{\sqrt{2\pi (1 - s^2)}} \exp \left( \frac{-sz_\ell - z_\varepsilon^2}{2(1 - s^2)} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) +$$

$$\int_{-z_\ell}^{z_\ell} s \Phi \left( \frac{z_\ell - sz_\ell}{\sqrt{1 - s^2}} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right).$$

(B.14)

The third term of the Equation B.14 is again a truncated normal with a scaling term of $\phi(z_\varepsilon)$ and the same $\vartheta$ and $\tilde{\sigma}$ as before. Hence,

$$\int_{-z_\ell}^{z_\ell} \frac{-s}{\sqrt{2\pi (1 - s^2)}} \exp \left( \frac{-(sz_\ell - z_\varepsilon^2)}{2(1 - s^2)} \right) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z_\ell^2}{2} \right) =$$

$$- s \phi(z_\varepsilon) \left( \left( \Phi \left( \frac{z_\ell - \vartheta}{\sqrt{1 - s^2}} \right) - \Phi \left( \frac{-z_\ell - \vartheta}{\sqrt{1 - s^2}} \right) \right) \vartheta -$$

148
Applying the law of iterated expectations to regions, we have that

\[
\sqrt{1 - s^2} \left( \phi \left( \frac{z_e - \bar{\theta}}{\sqrt{1 - s^2}} \right) - \phi \left( \frac{-z_e - \bar{\theta}}{\sqrt{1 - s^2}} \right) \right).
\]

The last term of the Equation B.14 is the product of the conditional normal distribution based on \( z_e \) and a normal distribution of \( z_e \). As a result, it is the probability of \( z_e \in [-z_e, z_e] \) and \( z_{e'} \leq z_e \) where \( z_e \) and \( z_{e'} \) are distributed according to

\[
\mathcal{N} \left( \begin{pmatrix} z_e \\ z_{e'} \end{pmatrix} \bigg| \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix} \right) .
\]

Lemma B.2.8. If \( \bar{E}_{z_e} \) denotes the quantity in Lemma B.2.6 and \( \bar{E}_{z_e,z_{e'}} \) denotes the quantity in Lemma B.2.7, then for \( z_e \neq z_{e'} \),

\[
\mathbb{E} (f_{z_e} f_{z_{e'}}) = \frac{1}{4}(P(z_e > z_e, z_{e'} > z_e) + P(z_e > -z_e, z_{e'} > -z_e)) + \frac{1}{2z_e} \bar{E}_{z_e} + \frac{1}{4z_e^2} \bar{E}_{z_e,z_{e'}}.
\]

Proof. We have that

\[
\mathbb{E} (f_{z_e} f_{z_{e'}}) = \mathbb{E} \left( \left( g(z_e) + \frac{1}{2} \right) \left( g(z_{e'}) + \frac{1}{2} \right) \mathbb{I} (z_e > -z_e, z_{e'} > -z_e) \right)
\]

\[
= \mathbb{E} (g(z_e)g(z_{e'})) \mathbb{I} (z_e > -z_e, z_{e'} > -z_e) + \frac{1}{4}P(z_e > -z_e, z_{e'} > -z_e).
\]

Applying the law of iterated expectations to regions, we have that

\[
\mathbb{E} (g(z_e)g(z_{e'})) = \mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e > z_e, z_{e'} > z_e \Bigg) P(z_e > z_e, z_{e'} > z_e) +
\]

\[
\mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e \in [-z_e, z_e], z_{e'} > z_e \Bigg) P(z_e \in [-z_e, z_e], z_{e'} > z_e) +
\]

\[
\mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e > z_e, z_{e'} \in [-z_e, z_e] \Bigg) P(z_e > z_e, z_{e'} \in [-z_e, z_e]) +
\]

\[
\mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e \in [-z_e, z_e], z_{e'} \in [-z_e, z_e] \Bigg) P(z_e \in [-z_e, z_e], z_{e'} \in [-z_e, z_e]).
\]

Because of Lemma B.2.6, we get that

\[
\frac{1}{4z_e} \bar{E}_{z_e} = \mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e \in [-z_e, z_e], z_{e'} > z_e \Bigg) P(z_e \in [-z_e, z_e], z_{e'} > z_e)
\]

and

\[
\frac{1}{4z_e} \bar{E}_{z_{e'}} = \mathbb{E} (g(z_e)g(z_{e'})) \bigg| z_e > z_e, z_{e'} \in [-z_e, z_e] \Bigg) P(z_e > z_e, z_{e'} \in [-z_e, z_e]).
\]

149
Similarly, Lemma B.2.7 gives us that

\[
\frac{1}{4z_e^2} \tilde{E}_{z_e, z_{e'}} = \mathbb{E} \left( g(z_e)g(z_{e'}) \right) \mid z_e \in [-z_e, z_e], z_{e'} \in [-z_{e'}, z_{e'}] \right) \mathcal{P}(z_e \in [-z_e, z_e], z_{e'} \in [-z_{e'}, z_{e'}]).
\]

If we plug those values in, we get the value stated in the lemma.

**Lemma B.2.9.** If \(Z_e, Z_{e'} \sim \mathcal{N}(0, \Sigma)\) and \(s = \text{Cov}(Z_e, Z_{e'})\) and \(z_e \in (0, 2]\), then

\[
-2\psi^{-1}(-z_e) \left( \Phi(-z_e) + \left( \frac{1}{2} - \Phi(-z_e) \right) \psi^{-1}(z_e) \right) \leq \mathbb{E} \left( \psi^{-1}(Z_e) \psi^{-1}(Z_{e'}) \right) - \mathbb{E} \left( f(Z_e)f(Z_{e'}) \right) \leq 0.
\]

**Proof.** To take advantage of the conditional Normal distribution, we can write

\[
\psi^{-1}(Z_e) \psi^{-1}(Z_{e'}) - f(Z_e)f(Z_{e'})
\]

as

\[
\frac{1}{2} \left( (\psi^{-1}(Z_e) - f(Z_e))(\psi^{-1}(Z_{e'}) + f(Z_{e'})) + (\psi^{-1}(Z_e) + f(Z_e))(\psi^{-1}(Z_{e'}) - f(Z_{e'})) \right).
\]

Without loss of generality, we focus on \(\mathbb{E} \left( \psi^{-1}(Z_e) - f(Z_e))(\psi^{-1}(Z_{e'}) + f(Z_{e'})) \right)\).

Given \(z_{e'}\), we can show that

\[
\int_0^\infty (\psi^{-1}(z_e) - f(z_e)) \phi \left( \frac{z_e - sz_{e'}}{\sqrt{1 - s^2}} \right) \, dz_e
\]

\[
= -\int_0^\infty (1 - \psi^{-1}(z_e)) - (1 - f(z_e))) \phi \left( \frac{z_e - sz_{e'}}{\sqrt{1 - s^2}} \right) \, dz_e
\]

\[
= -\int_0^\infty (\psi^{-1}(-z_e) - f(-z_e)) \phi \left( \frac{z_e - sz_{e'}}{\sqrt{1 - s^2}} \right) \, dz_e
\]

\[
= -\int_{-\infty}^0 (\psi^{-1}(z_e') - f(z_e')) \phi \left( \frac{-z_e' - sz_{e'}}{\sqrt{1 - s^2}} \right) \, dz_e'.
\]

Using this, we have that

\[
\mathbb{E} \left( \psi^{-1}(z_e) - f(z_e) \mid z_{e'} \right) = \int_{-\infty}^0 (\psi^{-1}(z_e) - f(z_e)) \left( \phi \left( \frac{z_e - sz_{e'}}{\sqrt{1 - s^2}} \right) - \phi \left( \frac{-z_e - sz_{e'}}{\sqrt{1 - s^2}} \right) \right) \, dz_e.
\]
We can also demonstrate that $\mathbb{E} (\psi^{-1}(z_{\ell}) - f(z_{\ell}) | z_{\ell'}) = -\mathbb{E} (\psi^{-1}(z_{\ell}) - f(z_{\ell}) | -z_{\ell'})$ because

$$
\phi \left( \frac{z_{\ell} - sz_{\ell'}}{\sqrt{1 - s^2}} \right) - \phi \left( \frac{-z_{\ell} - sz_{\ell'}}{\sqrt{1 - s^2}} \right) = - \left( \phi \left( \frac{z_{\ell} - s(-z_{\ell'})}{\sqrt{1 - s^2}} \right) - \phi \left( \frac{-z_{\ell} - s(-z_{\ell'})}{\sqrt{1 - s^2}} \right) \right).
$$

As a result, because $\psi^{-1}(z_{\ell}) - f(z_{\ell}) > 0$ for $z_{\ell} < 0$ and $\psi^{-1}(z_{\ell'}) + f(z_{\ell'}) < \psi^{-1}(-z_{\ell'}) + f(-z_{\ell'})$ for $z_{\ell'} < 0$, $\mathbb{E} (\psi^{-1}(Z_{\ell}) \psi^{-1}(Z_{\ell'})) - \mathbb{E} (f(Z_{\ell}) f(Z_{\ell'})) \leq 0$.

Now that we have established an upper bound, we need to establish a lower bound. From our discussion above, $\mathbb{E} (\psi^{-1}(Z_{\ell}) \psi^{-1}(Z_{\ell'})) - \mathbb{E} (f(Z_{\ell}) f(Z_{\ell'}))$ can be written as

$$
\int_{-\infty}^{0} \int_{-\infty}^{\infty} (\psi^{-1}(z_{\ell}) - f(z_{\ell})) (\psi^{-1}(z_{\ell'}) + f(z_{\ell'}) - (\psi^{-1}(-z_{\ell'}) + f(-z_{\ell'}))) \phi \left( \frac{z_{\ell} - sz_{\ell'}}{\sqrt{1 - s^2}} \right) \phi(z_{\ell'}) dz_{\ell} dz_{\ell'}.
$$

If we consider the parts, we have that

$$
-\psi^{-1}(-z_{\ell}) \leq \psi^{-1}(z_{\ell}) - f(z_{\ell}) \leq \psi^{-1}(-z_{\ell})
$$

and

$$
-2 \leq \psi^{-1}(z_{\ell'}) + f(z_{\ell'}) - (\psi^{-1}(-z_{\ell'}) + f(-z_{\ell'})) \leq -2 \psi^{-1}(z_{\ell}) \quad z_{\ell'} \in (-\infty, -z_{\ell})
$$

$$
-2 \psi^{-1}(z_{\ell}) \leq \psi^{-1}(z_{\ell'}) + f(z_{\ell'}) - (\psi^{-1}(-z_{\ell'}) + f(-z_{\ell'})) \leq 0 \quad z_{\ell'} \in [-z_{\ell}, 0].
$$

As a result, a lower bound is

$$
-2 \psi^{-1}(-z_{\ell}) \left( \Phi(-z_{\ell}) + \left( \frac{1}{2} - \Phi(-z_{\ell}) \right) \psi^{-1}(z_{\ell}) \right).
$$

Because a similar argument holds for $(\psi^{-1}(Z_{\ell}) + f(Z_{\ell})) (\psi^{-1}(Z_{\ell'}) - f(Z_{\ell'}))$, we get the bound described in the lemma.

Using these technical lemmas, we have the following lemma.

**Lemma B.2.10.** The circular correlation for the SvM-p-2 model described in (3.17) and under conditions of Lemma B.2.9 is

$$
\frac{2 \mathbb{E} \left( \sin(y_{\ell} - \alpha) \sin(y_{\ell'} - \alpha) \right)}{1 - \frac{1}{2} \left( \frac{I_2(\rho_1)}{I_0(\rho_1)} \cos(2(m_1 - \alpha)) + \frac{I_2(\rho_2)}{I_0(\rho_2)} \cos(2(m_2 - \alpha)) \right)},
$$

151
where $2\mathbb{E}(\sin(y_\ell - \alpha) \sin(y_{\ell'} - \alpha))$ is equal to

$$\frac{I_1(\rho_1) I_1(\rho_2)}{I_0(\rho_1) I_0(\rho_2)} (\cos(m_1 - m_2) - \cos(m_1 + m_2 - 2\alpha)) +$$

$$\mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right) \left( \left( \frac{I_1(\rho_1)}{I_0(\rho_1)} \right)^2 (1 - \cos(2(m_1 - \alpha))) + \right.$$  

$$\left( \frac{I_1(\rho_2)}{I_0(\rho_2)} \right)^2 (1 - \cos(2(m_2 - \alpha))) -$$  

$$2 \frac{I_1(\rho_1) I_1(\rho_2)}{I_0(\rho_1) I_0(\rho_2)} (\cos(m_1 - m_2) - \cos(m_1 + m_2 - 2\alpha)) \right),$$

and $\mathbb{E}(\psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}))$ is approximated in Lemma B.2.9.

**Proof.** For the denominator, according to B.2.4, $P(\zeta_\ell = 1) = P(\zeta_{\ell'} = 1) = \frac{1}{2}$. Further, $P(\zeta_\ell = 2) = 1 - P(\zeta_\ell = 1)$ and $P(\zeta_{\ell'} = 2) = 1 - P(\zeta_{\ell'} = 1)$ so $P(\zeta_\ell = 2) = P(\zeta_{\ell'} = 1) = \frac{1}{2}$ as well. Hence, we get the value in the denominator if we plug in these probabilities.

Meanwhile, for the numerator, we need to compute the probability for the four cases:

1. $\zeta_1, \zeta_2 = 1$
2. $\zeta_1, \zeta_2 = 2$
3. $\zeta_1 = 1, \zeta_2 = 2$
4. $\zeta_1 = 2, \zeta_2 = 1$.

For the first case, $P(\zeta_1 = 1, \zeta_2 = 1)$, which is approximated in Lemma B.2.9.

For the second, through our approximation,

$$P(\zeta_\ell = 2, \zeta_{\ell'} = 2) = \mathbb{E} \left( \mathbb{I}(\zeta_\ell, \zeta_{\ell'} = 2) \right)$$

$$= \mathbb{E} \left( \mathbb{E} \left( \mathbb{I}(\zeta_\ell, \zeta_{\ell'} = 2) \mid \lambda_1, \psi_2 \right) \right)$$

$$= \mathbb{E} \left( (1 - \lambda_1)(1 - \psi_2) \right)$$

$$= \mathbb{E} \left( (1 - \psi^{-1}(z_\ell))(1 - \psi^{-1}(z_{\ell'})) \right)$$

$$= 1 - \mathbb{E} \left( \psi^{-1}(z_\ell) \right) - \mathbb{E} \left( \psi^{-1}(z_{\ell'}) \right) + \mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right)$$

$$\approx \mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right).$$

As a result, it too is weighted by the approximation in Lemma B.2.9.
Finally, for the last two cases, the calculations are identical. Following the same calculation as above and using Lemma B.2.9, we get that

\[
P(\zeta_\ell = 1, \zeta_{\ell'} = 2) = \mathbb{E} \left( \psi^{-1}(z_\ell) \left( 1 - \psi^{-1}(z_{\ell'}) \right) \right) \\
= \mathbb{E} \left( \psi^{-1}(z_\ell) \right) - \mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right) \\
\approx \frac{1}{2} - \mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right).
\]

Similarly, \( P(\zeta_\ell = 2, \zeta_{\ell'} = 1) \) is also approximately equal to \( \frac{1}{2} - \mathbb{E} \left( \psi^{-1}(z_\ell) \psi^{-1}(z_{\ell'}) \right) \). We can again use Lemma B.2.4 to calculate this quantity.

If we plug these probabilities into the numerator, we get the quantity in the lemma.

We explicate at a high level the upper and lower bound. As shown above, \( \mathbb{E} \left( f(Z_\ell) f(Z_{\ell'}) \right) \) provides an upper bound on \( \mathbb{E} \left( \psi^{-1}(Z_\ell) \psi^{-1}(Z_{\ell'}) \right) \). This upper bound is comprised of computing the expectation over four areas: (1) \( f(Z_\ell) f(Z_{\ell'}) = 0 \), (2) \( f(Z_\ell) f(Z_{\ell'}) = 1 \), (3) either \( f(Z_\ell) = 1 \) or \( f(Z_{\ell'}) = 1 \), and (4) \( Z_\ell, Z_{\ell'} \in [-z_\epsilon, z_\epsilon] \times [-z_\epsilon, z_\epsilon] \). The technical lemmas discussed are for the calculation of \( \frac{1}{2z_\epsilon} \mathbb{E}_{z_\epsilon} \) and \( \frac{1}{4z_\epsilon^2} \mathbb{E}_{z_\epsilon,z_{\epsilon'}} \), which are the expectations from areas (3) and (4) respectively. In contrast, the lower bound without \( \mathbb{E} \left( f(Z_\ell) f(Z_{\ell'}) \right) \) is essentially the product of two terms. The first part, \( -\psi^{-1}(-z_\epsilon) \), represents the greatest amount our approximation can be larger than the inverse logit function. Doing so allows us to marginalize out one of the variables from the bivariate normal distribution. After some rearrangement, the other part again represents the greatest differences in particular regions multiplied by the probability of those regions. It can be more "refined" because we are back to considering one variable. However, this part also reflects the algebraic choices that allowed us to transform the difference of products to a sum of products of terms that only depend on \( Z_\ell \) or \( Z_{\ell'} \). Because this bound in Lemma B.2.9 gets smaller as \( z_\epsilon \) increases, the bound is minimized at \( z_\epsilon = 2 \).

### B.3 MCMC sampling

To sample from our models, we turn to Hamiltonian Monte Carlo (HMC). In HMC, each parameter that we are interested in sampling is given a momentum random variable [51]. If we use the notation from Neal’s paper to introduce HMC, a parameter is represented by \( q_i \) and its momentum is represented by \( p_i \). Instead of sampling from the posterior, we are now interested in sampling the joint distribution based on the Hamiltonian, i.e. \( \frac{1}{T} \exp \left( -\frac{U(q) + K(p)}{T} \right) \). For this distribution based on the Hamiltonian, the potential energy, \( U(q) \), is defined to be the negative log posterior. On the other hand, the momentum, \( K(p) \), is defined to be \( p^T M p \) for some mass matrix \( M \). If we use the
leapfrog integrator, the sampling updates are the following:

\[
p_i(t + \epsilon) = p_i(t) - \epsilon \frac{\partial U}{\partial q_i}(q(t)),
q_i(t + \epsilon) = q_i(t) + \epsilon(M^{-1}p(t + \frac{\epsilon}{2}))_i,
p_i(t + \epsilon) = p_i(t + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \frac{\partial U}{\partial q_i}(q(t + \epsilon)).
\]

To understand why we need these different approaches, we will derive the updates for the momentum vector to better understand how HMC will fit our models because this update is the derivative of the log posterior with respect to the current parameters \[51\]. This is one of the key interactions between the sampler and the model. In particular, we will focus on how the momentum terms related to the Gaussian Process term will be updated. While this doesn’t give us a complete picture of the HMC sampler, we do so because the updates for the other parameters are relatively straightforward. It will also illustrate the challenges of fitting these models.

### B.3.1 SvM and SvM-c models

**SvM model** As discussed earlier, \( m \) is distributed according to a Projected Gaussian Process. In their paper, Wang and Gelfand suggest sampling an auxiliary variable \( r \in \mathbb{R}^N \) such that if \( Z_1 \) and \( Z_2 \) are distributed according to the description in (B.3), \( z_{1,\ell} = r_\ell \cos(m_\ell) \) and \( z_{2,\ell} = r_\ell \sin(m_\ell) \) for \( \ell \in 1, 2, \ldots, N \) \[71\]. In doing so, we have that

\[
p(r, m) = \left( \prod_{\ell=1}^N r_\ell \right) \mathcal{N}\left( \begin{pmatrix} r \cos(m), r \sin(m) \end{pmatrix} \middle| (\mu_1, \mu_2), I_2 \otimes \Sigma \right). \tag{B.15}
\]

If we refer to the parametrization in (B.15) as the *circular centered parametrization*, the updates for the momentum vector corresponding to \( m_\ell \) and \( r_\ell \) are respectively

\[
-\frac{\partial U}{\partial q_{m_\ell}}(q(t)) = \left( \rho_\ell \cos(y_\ell) + r_\ell \left( \Sigma^{-1} \begin{pmatrix} \cos(m) - \mu_1 \end{pmatrix} \right) \right) \sin(m_\ell) - \\
\left( \rho_\ell \sin(y_\ell) + r_\ell \left( \Sigma^{-1} \begin{pmatrix} \sin(m) - \mu_2 \end{pmatrix} \right) \right) \cos(m_\ell), \tag{B.16}
\]

\[
-\frac{\partial U}{\partial q_{r_\ell}}(q(t)) = \frac{1}{r_\ell} - \left( \Sigma^{-1} \begin{pmatrix} \cos(m) - \mu_1 \end{pmatrix} \right) \cos(m_\ell) - \\
\left( \Sigma^{-1} \begin{pmatrix} \sin(m) - \mu_1 \end{pmatrix} \right) \sin(m_\ell). \tag{B.17}
\]

We do not necessarily have to transform \( r \) and \( m \) directly to draws from the bivariate Gaussian Process. Instead, we can first transform them to draws from \( 2N \) independent \( \mathcal{N}(0, 1) \) distribution.
In other words, if $(\tilde{Z}_1, \tilde{Z}_2) \sim \mathcal{N}(0, \mathbb{I}_2 \otimes \mathbb{I}_N)$, let $\tilde{z}_{1,\ell} = \tilde{r}_\ell \cos(\tilde{m}_\ell)$ and $\tilde{z}_{2,\ell} = \tilde{r}_\ell \sin(\tilde{m}_\ell)$ for $\tilde{r} \in \mathbb{R}^{N+}$, $\tilde{m} \in [0, 2\pi)^N$, and $\ell = 1, 2, \ldots, N$. Then, in a simplification of (B.15), we have that for $\ell = 1, 2, \ldots, N$, 

$$p(\tilde{r}_\ell) = r_\ell \exp\left\{-\frac{1}{2} r_\ell^2\right\} \quad p(\tilde{m}_\ell) = \text{Unif}(0, 2\pi). \quad (B.18)$$

Let $L$ denote the Cholesky decomposition of the covariance matrix of the Gaussian Process, $\Sigma$. We can then set $Z_1 = L\tilde{Z}_1 + \mu_1$ and $Z_2 = L\tilde{Z}_2 + \mu_2$ to change these draws from independent normal distributions to draws from our bivariate Gaussian Process. Finally, we can use the arctan* defined in 3.1 to transform $Z_1$ and $Z_2$ to $m$.

If we refer to the parametrization in (B.18) as the circular non-centered parametrization and use notation from before, the updates for the momentum vector corresponding to $\tilde{m}_\ell$ and $\tilde{r}_\ell$ are respectively

$$- \frac{\partial U}{\partial q_{\tilde{m}_\ell}}(q(t)) = \sum_{\ell'=1}^{N} -\nu_\ell \sin(y_{\ell'} - m_{\ell'}) \left(\frac{z_{2,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \tilde{r}_\ell \sin(\tilde{m}_\ell) + \frac{z_{1,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \tilde{r}_\ell \cos(\tilde{m}_\ell)\right), \quad (B.19)$$

$$- \frac{\partial U}{\partial q_{r_\ell}}(q(t)) = \sum_{\ell'=1}^{N} -\nu_\ell \sin(y_{\ell'} - m_{\ell'}) \left(-\frac{z_{2,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \cos(\tilde{m}_\ell) + \frac{z_{1,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \sin(\tilde{m}_\ell)\right) + \frac{1}{r_\ell} - \tilde{r}_\ell. \quad (B.20)$$

We will now compare the updates in (B.16) and (B.17) to the updates in (B.19) and (B.20). We immediately see that the centered updates are much more expensive because they involve inverting the covariance matrix. If we look closer and set the momentum updates for $m_\ell$ and $\tilde{m}_\ell$ to be 0, we see that

$$\frac{\sin(m_\ell)}{\cos(m_\ell)} = \frac{\left(\nu_\ell \sin(y_{\ell'}) + r_\ell \left(\Sigma^{-1}(r \sin(m) - \mu_2)\right)_{\ell'}\right)}{\left(\nu_\ell \cos(y_{\ell'}) + r_\ell \left(\Sigma^{-1}(r \cos(m) - \mu_1)\right)_{\ell'}\right)},$$

$$\frac{\sin(\tilde{m}_\ell)}{\cos(\tilde{m}_\ell)} = -\sum_{\ell'=1}^{N} \nu_\ell \sin(y_{\ell'} - m_{\ell'}) \frac{z_{1,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \tilde{r}_\ell \left(\sum_{\ell'=1}^{N} \nu_\ell \sin(y_{\ell'} - m_{\ell'}) \frac{z_{1,\ell'}}{z_{1,\ell'}^2 + z_{2,\ell'}^2} L_{\ell',\ell} \tilde{r}_\ell\right).$$

From this, we see that the update for $m_\ell$ tries to balance information from the data weighted by the concentration parameter $\nu_\ell$ against the prior information weighted by $r_\ell$. However, how the prior
information is used is not fully clear because the difference between the current positions and the prior means are multiplied by the inverse of the covariance matrix. On the other hand, the update for \( \tilde{m}_\ell \) separates the terms dependent on \( \sin(\tilde{m}_\ell) \) and \( \cos(\tilde{m}_\ell) \). Of note, these terms are weighted by the concentration parameter at the location, the Cholesky decomposition, and \( r_\ell \). We see that the ratio is reversed and negated compared to the update for \( m_\ell \). The derivative of \( \arctan^\ast(z_{1,\ell}, z_{2,\ell}) \) might cause this because as seen in Lemma B.4.1 in the Appendix, it assigns information from the other coordinates.

On the other hand, the update for \( r_\ell \) does not explicitly involve the data whereas the update for \( e_{r\ell} \) does. This update makes it clear that \( r_\ell \) is an auxiliary variable. This might be an issue because as discussed earlier, \( r_\ell \) affects how strongly to weigh prior information and it might be good for the data to affect such a term. The update still does provide some control due to the \( \frac{1}{r_\ell^3} \) term, which reduces the gradient for large values. Unfortunately, this may be counteracted by prior information and/or the current position of \( m_\ell \). Meanwhile, the update for \( \tilde{r}_\ell \) involves both \( -\tilde{r}_\ell \) and \( \frac{1}{r_\ell^3} \). This will help discourage both large and small values. Further, all observations weigh in unless as discussed previously, the ratio of \( \frac{\sin(\tilde{m}_\ell)}{\cos(\tilde{m}_\ell)} \) matches the ratio of the information from the data.

There is another additional fundamental issue with the non-centered circular parametrization. We sample \( r \) and \( m \) only because this gives us control over \( m \) while providing an easier space to sample from. When we sample \( \tilde{r} \) and \( \tilde{m} \) as an analogue to this approach and transform it to \( m \), this transformation is taking a convex combination of \( \tilde{r} \cos(\tilde{m}) \) and \( \tilde{r} \sin(\tilde{m}) \) with the corresponding Cholesky decomposition row as weights. It is as if we are sampling \( m \) through \( Z_1 \) and \( Z_2 \). This becomes problematic because linear moves in \( Z_1 \) and \( Z_2 \) do not translate to linear moves in \( m \). For instance, a linear move along a ray from the origin does not change \( m \).

As a result, we turn to elliptical slice sampling [50].

**SVM-c** For this model, we have to marginalize out the cluster membership in order to sample for \( z_{\ell} \). In other words, the probability for an observation \( y_\ell \) is the following.

\[
p(y_\ell | m_k, \rho_k, \lambda) = \sum_k \lambda_k f_V(y_\ell | m_k, \rho_k, \lambda).
\]

The updates are slightly altered. The centered update for the momentum vector for the SVM-c described in (3.15) corresponding to \( z_{k,\ell} \) becomes

\[
- \frac{\partial U}{\partial q_{k,1,\ell}} q_{k,1,\ell}(t) = \frac{\exp(\rho_{k,\ell} \cos(y_\ell - m_k, \ell))}{\sum_k \lambda_k f_V(y_\ell | m_k, \rho_k, \ell)} \left( -\frac{z_{k,2,\ell}}{z_{k,1,\ell}^2 + z_{k,2,\ell}^2} \rho_{k,\ell} \sin(y_\ell - m_k, \ell) \right) - (\Sigma_k^{-1} z_k)\ell \tag{B.21}
\]

\[
- \frac{\partial U}{\partial q_{k,2,\ell}} q_{k,2,\ell}(t) =
\]

156
Alternatively, if we let \( \Sigma = \Sigma_k \), the non-centered update changes to

\[
\frac{\exp(\rho_{k,\ell} \cos(y_\ell - m_{k,\ell}))}{\sum_k \lambda_k f_{Y}(y_\ell \mid m_{k,\ell}, \rho_{k,\ell})}\left(\frac{z_{k,1,\ell}}{z_{k,1,\ell}^2 + z_{k,2,\ell}^2} \rho_{k,\ell} \sin(y_\ell - m_{k,\ell})\right) - (\Sigma_k^{-1} z_k)_\ell \quad (B.22)
\]

The non-centered update changes to

\[
- \frac{\partial U}{\partial q_{k,1,\ell}} q_{k,1,\ell}(t) = - \frac{\partial U}{\partial q_{k,2,\ell}} q_{k,2,\ell}(t) = - \frac{\partial}{\partial q_{k,1,\ell}} q_{k,1,\ell}(t)
\]

\[
= \sum_{\ell=1}^{N} \left[ \frac{\exp(\rho_{k,\ell} \cos(y_\ell - m_{k,\ell}))}{\sum_k \lambda_k f_{Y}(y_\ell \mid m_{k,\ell}, \rho_{k,\ell})}\left(\frac{-z_{k,2,\ell}}{z_{k,1,\ell}^2 + z_{k,2,\ell}^2} \rho_{k,\ell} \sin(y_\ell - m_{k,\ell})\right) \right] L_\ell \cdot \bar{z}_{k,\ell}, \quad (B.23)
\]

\[
= \sum_{\ell=1}^{N} \left[ \frac{\exp(\rho_{k,\ell} \cos(y_\ell - m_{k,\ell}))}{\sum_k \lambda_k f_{Y}(y_\ell \mid m_{k,\ell}, \rho_{k,\ell})}\left(\frac{z_{k,1,\ell}}{z_{k,1,\ell}^2 + z_{k,2,\ell}^2} \rho_{k,\ell} \sin(y_\ell - m_{k,\ell})\right) \right] L_\ell \cdot \bar{z}_{k,\ell}. \quad (B.24)
\]

These updates are similar to the \( \text{SvM} \) updates from B.16, B.17, B.19, and B.20. The difference is that these updates are also weighted by how much an observation belongs to that particular’s component. Because these distinctions aren’t substantial, we again will use the elliptical slice sampling approach to sample for \( (Z_{k,1} - \mu_{k,1}, Z_{k,2} - \mu_{k,2}) \mid \varphi_k, \nu_k, \zeta \) for \( k \in 1, 2, \ldots K \). The only modification from the proposal in the previous section is that the likelihood is now defined to be the following:

\[
L(y \mid m_k, \varphi_k, \nu_k, \zeta) = \prod_{\ell} (f_{Y}(y_\ell \mid m_\ell, \rho_\ell))^\Pi(\zeta=\ell).
\]

### B.3.2 SvM-p model

Here, the calculations are much more straightforward. Because the Gaussian Process assigns probability through a multivariate Normal distribution, there are two ways to sample \( z_\ell \) without loss of generality. We can sample \( z_\ell \) directly from \( \text{GP}(0, \Sigma) \) or the so-called centered parametrization. Alternatively, if we let \( \Sigma = LL^T \) and \( z = L\bar{z} \), we can sample \( z_\ell \) according to the non-centered parametrization. We will present the updates from both parametrizations.

We again marginalize out the cluster assignment to sample for \( z_\ell \). The probability for a single observation \( y_\ell \) is the following:

\[
p(y_\ell \mid z_\ell, m, \rho) = \sum_k \lambda_k f_{Y}(y_\ell \mid m_k, \rho_k).
\]

Then, the centered parametrization update for the model described in (3.16) is

\[
- \frac{\partial U}{\partial q_{k,\ell}}(q(t)) = \lambda_{k,\ell} f_{Y}(y_\ell \mid m_k, \rho_k) - \sum_{k'=1}^{K} \lambda_{k',\ell} f_{Y}(y_\ell \mid m_{k'}, \rho_{k'}) \frac{p(y_\ell \mid z_\ell, m, \rho)}{p(y_\ell \mid z_\ell, m, \rho)} - (\Sigma^{-1} z)_\ell. \quad (B.25)
\]
In the two components case, this simplifies to
\[-\frac{\partial U}{\partial q_{k,\ell}}(q(t)) = \lambda_{1,\ell}(1 - \lambda_{1,\ell}) \frac{f_V(y_{\ell} | m_1, \rho_1) - f_V(y_{\ell} | m_2, \rho_2)}{p(y_{\ell} | z_{\ell}, m_1, m_2, \rho_1, \rho_2)} - (\Sigma^{-1}z_\ell).\] (B.26)

The non-centered parametrization update is
\[-\frac{\partial U}{\partial q_{\ell}}(q(t)) = \sum_{\ell'} L_{\ell',\ell} \lambda_{1,\ell'} \frac{f_V(y_{\ell'} | m_{k,\ell'}) - \sum_{k'=1}^K \lambda_{k',\ell'} f_V(y_{\ell'} | m_{k',\ell'})}{p(y_{\ell'} | z_{\ell'}, m, \rho)} - \bar{z}_{\ell}.\] (B.27)

The non-centered parametrization update in the two components case reduces to
\[-\frac{\partial U}{\partial q_{\ell}}(q(t)) = \sum_{\ell'} \lambda_{1,\ell'}(1 - \lambda_{1,\ell'}) \frac{f_V(y_{\ell'} | m_1, \rho_1) - f_V(y_{\ell'} | m_2, \rho_2)}{p(y_{\ell'} | z_{\ell'}, m_1, m_2, \rho_1, \rho_2)} L_{\ell',\ell} - \bar{z}_{\ell}.\] (B.28)

From doing so, we see that the centered and non-centered parametrization differ in how they use the likelihood for updates and how they use covariance information. The centered update based on the likelihood compares the probability of an observation belonging to component \(k\) against the probability of the observation belonging to the current mixture of von Mises distributions. This update is also weighted by the currently probability of the observation belongs to component \(k\). In the two component case, this becomes a comparison the probability of the observation belonging to either component. On the other hand, the non-centered update uses the covariance information to pool these weighted comparisons. Because these updates seem sensible, we use HMC to sample from \(SvM-p\).

### B.3.3 Regularized expectation maximization

To help with the sampling, we initialized it with results from applying Expectation Maximization (EM) to the unnormalized posterior or a regularized EM. Because \(\Sigma\) might be close to singular, we ran regularized EM on the non-centered parametrization for our updates.

**SvM-c** Within the regularized EM framework, we let \(\zeta_{\ell}\) for \(\ell = 1, 2, \ldots, N\) be the latent variable and \(\lambda_k, \varphi_k, \nu_k, \text{ and } \bar{z}_k\) for \(k = 1, 2, \ldots, K\) be the parameters. If we summarize the parameters as \(\Theta\), the posterior can be written as following.

\[p(\zeta, \Theta | y, x) := \prod_{\ell} \prod_k (\lambda_k f_V(y_{\ell} | m_k, \rho_k))^{\mathbb{1}(\zeta_{\ell}=k)} \prod_k N(\bar{z} | 0, \mathbb{I}_n) \prod_{\ell} \prod_k N(\varphi_{k,\ell} | \nu_k, \varsigma^2) \prod_k N(\nu_k | 0, \tau^2).\]

For the expected conditional log unnormalized posterior, we need a term to represent \(P(\zeta_{\ell} =
\( k \mid \mathbf{y}, \lambda_k, \mathbf{z}_k, \varphi, \mathbf{t} \). We let \( r_{k,\ell} \) be this term. To be explicit, we are attaching the draws from the Gaussian Process, \( \mathbf{z}_k \), to \( m_{\cdot,\cdot} \). Then, it is defined in the following manner.

\[
r_{k,\ell} := \frac{\lambda_k f_V (y_\ell \mid m_{k,\ell}, \rho_{k,\ell})}{\sum_{k'} \lambda_{k'} f_V (y_\ell \mid m_{k',\ell}, \rho_{k',\ell})}.
\]  

With this term, the expected conditional log unnormalized posterior is the following.

\[
\mathbb{E} \left( \log (p(\zeta, \Theta \mid \mathbf{y}, \mathbf{x})) \right) = \sum_\ell \sum_k r_{k,\ell} (\log (\lambda_k) + \log (f_V (y_\ell \mid m_{k,\ell}, \rho_{k,\ell}))) + \\
\sum_\ell \sum_k \log (N(\mathbf{z}_k, \cdot | 0, 1)) + \\
\sum_\ell \sum_k \log (N(\varphi_{k,\ell} \mid \nu_k, \varsigma^2)) + \sum_k \log (N(\nu_k \mid 0, \tau^2)).
\]  

With this, we can now state the regularized EM algorithm.

- **E step:** Calculate the expected conditional log unnormalized posterior from B.30 and \( r_{k,\ell} \) from B.29.

- **M step:**
  - For \( k = 1, 2, \ldots, K \), set \( \lambda_k = \frac{1}{N} \sum_{\ell} r_{k,\ell} \).
  - For \( k = 1, 2, \ldots, K \), alternate updating \( \mathbf{z}_{k,1.} \) and \( \mathbf{z}_{k,2.} \) using coordinate gradient ascent and the following gradients.
    \[
    \frac{\partial}{\partial \mathbf{z}_{k,1,\ell}} \mathbb{E} \left( \log (p(\zeta, \Theta \mid \mathbf{y}, \mathbf{x})) \right) = \\
    \sum_{\ell'} L_{\ell',\ell} r_{k,\ell'} \frac{-z_{k,2,\ell'}}{z_{k,1,\ell'}^2 + z_{k,2,\ell'}^2} \rho_{k,\ell'} \sin (y_{\ell'} - m_{k,\ell'}) - \mathbf{z}_{k,\ell}.
    \]
    \[
    \frac{\partial}{\partial \mathbf{z}_{k,2,\ell}} \mathbb{E} \left( \log (p(\zeta, \Theta \mid \mathbf{y}, \mathbf{x})) \right) = \\
    \sum_{\ell'} L_{\ell',\ell} r_{k,\ell'} \frac{z_{k,1,\ell'}}{z_{k,1,\ell'}^2 + z_{k,2,\ell'}^2} \rho_{k,\ell'} \sin (y_{\ell'} - m_{k,\ell'}) - \mathbf{z}_{k,\ell}.
    \]
  - For \( k = 1, 2, \ldots, K \) and \( \ell = 1, 2, \ldots N \), update \( \varphi_{k,\ell} \) using gradient ascent and the gradient below.
    \[
    \frac{\partial}{\partial \varphi_{k,\ell}} \mathbb{E} \left( \log (p(\zeta, \Theta \mid \mathbf{y}, \mathbf{x})) \right) = \\
r_{k,\ell} \rho_{k,\ell} \left( \cos (y_{\ell} - m_{k,\ell}) - \frac{I_{-1}(\rho_{k,\ell})}{I_0(\rho_{k,\ell})} \right) - \frac{1}{\varsigma^2} (\varphi_{k,\ell} - \nu_k).
    \]}
For \( k = 1, 2, \ldots, K \), set \( \nu_k = \frac{\sum_{\ell} r_{k,\ell} \phi_{k,\ell} \varsigma^2}{\varsigma^2 + 1} \).

We wish to discuss the updates in further detail. The updates for \( \lambda_k \) and \( \nu_k \) are similar to the maximum likelihood estimator for a multinomial distribution’s probabilities and a posterior conjugate Normal distribution’s mean. However, because we don’t know the assignments, we weight each observation with \( r_{k,\ell} \) instead. Next, to understand the update for \( \phi_{k,\ell} \), we note that according to the definition of the modified Bessel function in 3.3, \( I_{-1}(\rho_k) = I_1(\rho_k) \) or the expected concentration for a von Mises distribution. The update is intuitively trying to find the concentration parameter that best matches the expected concentration of that observation to the observation’s sample concentration weighted by how much the observations belongs to that cluster. The hierarchical prior helps the update because the update is also trying to match the concentration parameter against the hierarchical mean, \( \nu_k \), and can borrow strength from other observations.

Even though we did not derive the Hamiltonian Monte Carlo SvM non-centered parametrization momentum update for \( \tilde{z}_{k,\ell} \), that update is similar to the regularized EM update for \( \tilde{z}_{k,\ell} \). At a high level, this makes sense because \( \zeta_\ell \) assigns an observation to one component. Since we can only compute the conditional probability of \( \zeta_\ell \), we have to weight the update by \( r_{k,\ell} \). It is still interesting that we can use this standard parametrization. It might be possible to do so because we are only interested in optimizing a point estimate and not interested in exploring the space. One consequence of this similarity is that like Hamiltonian Monte Carlo, we choose to use a coordinate gradient ascent approach to update all \( \tilde{z}_k \) because an update for a coordinate can be written as the matrix multiplication of the transpose of the Cholesky and the other gradient information.

**SvM-p** Within the regularized EM framework, we let \( \zeta_\ell \) for \( \ell = 1, 2, \ldots, N \) be the latent variable and \( \tilde{z}_k, m_k, \) and \( \rho_k \) for \( k = 1, 2, \ldots, K \) be the parameters. To be clear, we attach the draw from the Gaussian Process, \( \tilde{z}_k \), to \( \lambda_k \). If we again summarize the parameters as \( \Theta \), the unnormalized posterior can be written as following.

\[
\bar{p}(\zeta, \Theta | y, x) := \prod_{\ell} \prod_k (\lambda_{k,\ell} f_V(y_\ell | m_k, \rho_k))^{\mathbb{I}(\zeta_\ell = k)} \prod_k N(\tilde{z}_k | 0, \mathbb{I}_N) \prod_k \Gamma(\rho_k | 1, 1) \prod_k \text{Unif}(m_k | 0, 2\pi).
\]

For the expected conditional log likelihood, we need to tweak our definition of \( r_{k,\ell} \) to remove the observation index from the von Mises distribution parameters and put them on the mixing probability. In other words, we let \( r_{k,\ell} \) be the following.

\[
r_{k,\ell} := \frac{\lambda_{k,\ell} f_V(y_\ell | m_k, \rho_k)}{\sum_{k'} \lambda_{k',\ell} f_V(y_\ell | m_k', \rho_k')}, \quad (B.31)
\]
Then, the expected conditional log unnormalized posterior is the following:

$$
\mathbb{E}(\log(\tilde{p}(\zeta, \Theta | y, x))) = \sum_{\ell} \sum_{k} r_{k, \ell} (\log(\lambda_{k, \ell}) + \log(f_V(y_{\ell} | m_k, \rho_k)) + \sum_{k} \log(N(z_k | 0, I_N)) + \sum_{k} \log(\text{Unif}(m_k | 0, 2\pi)) + \sum_{k} \log(\Gamma(\rho_k | 1, 1)).
$$

(B.32)

Our regularized EM algorithm is the following.

- **E step**: Calculate the expected conditional log unnormalized posterior from B.32 and \(r_{k, \ell}\) from B.31.

- **M step**:
  - For \(k = 1, 2, \ldots, K - 1\), update \(\tilde{z}_k\) using coordinate gradient ascent and the following gradient.
    $$
    \frac{\partial}{\partial \tilde{z}_{k, \ell}} \mathbb{E}(\log(\tilde{p}(\zeta, \Theta | y, x))) = \sum_{\ell'} (r_{k, \ell'} - \lambda_{k, \ell'}) L_{\ell', \ell} - \tilde{z}_{k, \ell}.
    $$
  - For \(k = 1, 2, \ldots, K\), set \(m_k = \arctan^*(\frac{\sum_{\ell} r_{k, \ell} \sin(y_{\ell})}{\sum_{\ell} r_{k, \ell} \cos(y_{\ell})})\).
  - For \(k = 1, 2, \ldots, K\), update \(\rho_k\) using gradient ascent and the following gradient.
    $$
    \frac{\partial}{\partial \rho_k} \mathbb{E}(\log(\tilde{p}(\zeta, \Theta | y, x))) = \sum_{\ell=1}^{n} r_{k, \ell} \left( \cos(y_{\ell} - m_k) - \frac{I_{-1}(\rho_k)}{I_0(\rho_k)} \right) - 1.
    $$

Note that the order of means can be kept by sorting the results at the end. Then, we make a few remarks about these updates. First, the updates for \(m_k\) and \(\rho_k\) in the SvM-\(p\) model are simpler than the updates in the SvM-c model due to our model choices. For instance, because the prior on \(m_k\) does not depend on \(m_k\), \(m_k\) is maximized when it is set equal to the weighted circular mean. The update becomes slightly more complicated if we use a von Mises distribution as a prior instead. Next, the update for \(\rho_k\) reflects the fact that the best \(\rho_k\) for all the weighted observation makes the expected concentration match the weighted sample concentration. This update does not consider the effect of the prior. It seems that in this case, the prior is providing a constant pull back to 0. This constant pull back is controlled by the \(b\) parameter in the prior Gamma distribution. If \(a\) is not 1, then the gradient will also include a \(\frac{1}{\rho_k}\) term.

**iVM** While this is an inversion of how we introduced the models, we do so because we can collect the parts independent of location information from the EM algorithm for the previous two
models for the iVM model’s regularized EM algorithm. As before, we let $\zeta_\ell$ for $\ell = 1, 2, \ldots, N$ be the latent variable and $\lambda_k, m_k, \rho_k$ for $k = 1, 2, \ldots, K$ be the parameters. If we again summarize the parameters as $\Theta$, the unnormalized posterior can be written as following.

$$
\tilde{p}(\zeta, \Theta \mid y, x) := \prod_\ell \prod_k (\lambda_k f_V(y_\ell \mid m_k, \rho_k))^{\Pi(\zeta_\ell = k)} \prod_k \Gamma(\rho_k \mid a_k, b_k) \prod f_V(m_k \mid u_k, c_k).
$$

We again need to tweak the definition of $r_{k,\ell}$. In other words, we let $r_{k,\ell}$ be the following.

$$
r_{k,\ell} := \frac{\lambda_k f_V(y_\ell \mid m_k, \rho_k)}{\sum_{k'} \lambda_{k'} f_V(y_\ell \mid m_{k'}, \rho_{k'})}.
$$  \hspace{1cm} (B.33)

Then the expected conditional log unnormalized posterior is the following:

$$
\mathbb{E} \left( \log(\tilde{p}(\zeta, \Theta \mid y, x)) \right) = \sum_{\ell} \sum_k r_{k,\ell} \left( \log(\lambda_k) + \log(f_V(y_\ell \mid m_k, \rho_k)) \right) + \sum_k \log(f_V(m_k \mid u_k, c_k)) + \sum_k \log(\Gamma(\rho_k \mid a_k, b_k)). \hspace{1cm} (B.34)
$$

For simplicity, we assume that $c_k = 0$ and $a_k = 1$ and $b_k = 0$. At a high level, we have an uniform prior on the mean direction and have a flat prior on the concentration parameter. While the latter is an improper prior, the posterior is still proper. Our EM algorithm is the following.

- **E step:** Calculate the expected conditional log unnormalized posterior from B.34 and $r_{k,\ell}$ from B.33.

- **M step:**
  - For $k = 1, 2, \ldots, K$, set $\lambda_k = \frac{1}{N} \sum_\ell r_{k,\ell}$.
  - For $k = 1, 2, \ldots, K$, set $m_k = \arctan^{*} \left( \frac{\sum_\ell r_{k,\ell} \sin(y_\ell)}{\sum_\ell r_{k,\ell} \cos(y_\ell)} \right)$.
  - For $k = 1, 2, \ldots, K$, update $\rho_k$ using gradient ascent and the following gradient.

$$
\frac{\partial}{\partial \rho_k} \mathbb{E} \left( \log(\tilde{p}(\zeta, \Theta \mid y, x)) \right) = \sum_{\ell=1}^n r_{k,\ell} \left( \cos(y_\ell - m_k) - \frac{I_{-1}(\rho_k)}{I_0(\rho_k)} \right).
$$

**B.3.4 Calculating posterior predictive probability**

Let $x^*$ represent the withheld locations, $y^*$ the withheld data, $\theta$ a posterior draw for the parameters based on $x$ and $y$, and $\theta^*$ a draw for the parameters for $x^*$ and $y^*$. Set $N$ to be the number of locations and $N^*$ to be the number of withheld locations. The posterior predictive probability is
We now discuss how to calculate it for our models.

**SvM-c:** Here, \( \theta = \{ \{ z_{k,1}, z_{k,2} \}_{k=1}^K, \{ \varphi_k \}_{k=1}^K, \{ \nu_k \}_{k=1}^K, \lambda_{k=1}^K \} \). Further, for the two Gaussian processes of component \( k \), let \( \Sigma_{k,1}^* \) and \( \Sigma_{k,2}^* \) be the covariance matrices based on \( x^* \) and \( \tilde{\Sigma}_{k,1} \) and \( \tilde{\Sigma}_{k,2} \) be the covariance matrices based on \( x \) and \( x^* \). In other words, \( \tilde{\Sigma}_{k,1,i,j} = K(x_i, x_j^*) \) for some kernel function \( K(\cdot, \cdot) \). Then, the posterior predictive probability for posterior draw \( i \) can be found in the following procedure:

1. Draw \( z_{k,1}^* \sim N(\mu_{k,1} + \tilde{\Sigma}_{k,1}^T \Sigma_{k,1}^{-1}(z_{k,1} - \mu_{k,1}), \Sigma_{k,1}^* - \tilde{\Sigma}_{k,1}^T \Sigma_{k,1}^{-1} \tilde{\Sigma}_{k,1}) \) and \( z_{k,2}^* \sim N(\mu_{k,2} + \tilde{\Sigma}_{k,2}^T \Sigma_{k,2}^{-2}(z_{k,2} - \mu_{k,2}), \Sigma_{k,2}^* - \tilde{\Sigma}_{k,2}^T \Sigma_{k,2}^{-2} \tilde{\Sigma}_{k,2}) \) for \( k = 1, 2, \ldots, K \).

2. Set \( m_k^* = \arctan^*(z_{k,1}^*, z_{k,2}^*) \).

3. Draw \( \varphi_k^* \sim N(\nu_k, \sigma^2) \). Set \( \rho^* = \exp(\varphi_k^*) \).

4. Let \( p_{i,1}^* = \prod_{n=1}^{N^*} (\sum_k \lambda_k f_N(y_{n}^* | m_{k,n}^*, \rho_{k,n}^*)) \).

5. Repeat the previous steps \( M - 1 \) more times for some \( M \in \mathbb{N} \) and denote the \( m^{th} \) result as \( p_{i,m}^* \).

6. If there are \( I \) iterations, return \( \frac{1}{M} \sum_m \frac{1}{I} \sum_i p_{i,m}^* \) for the posterior predictive probability.

**SvM-p:** Here, \( \theta = \{ \{ z_{k,1}, z_{k,2} \}_{k=1}^K, \{ m_k \}_{k=1}^K, \{ \rho_k \}_{k=1}^K \} \). Further, for the Gaussian processes of component \( k \), let \( \Sigma_k^* \) be the covariance matrix based on \( x^* \) and \( \tilde{\Sigma}_k \) be the covariance matrices based on \( x \) and \( x^* \). In other words, \( \tilde{\Sigma}_{k,i,j} = K(x_i, x_j^*) \) for some kernel function \( K(\cdot, \cdot) \). Then, the posterior predictive probability for posterior draw \( i \) can be found in the following procedure:

1. Draw \( z_k^* \sim N(\mu_k + \tilde{\Sigma}_k^T \Sigma_k^{-1}(z_k - \mu_k), \Sigma_k^* - \tilde{\Sigma}_k^T \Sigma_k^{-1} \tilde{\Sigma}_k) \) for \( k = 1, 2, \ldots, K - 1 \).

2. For \( n^* \in 1, 2, \ldots, N^* \), set \( \lambda_{n,n^*}^* = \Psi^{-1}(z_{1,n^*}, z_{2,n^*}, \ldots, z_{K-1,n^*}) \).

3. Let \( p_{i,1}^* = \prod_{n^*} (\sum_k \lambda_{k,n^*} f_N(y_{n^*} | m_k, \rho_k)) \).

4. Repeat the previous steps \( M - 1 \) more times for some \( M \in \mathbb{N} \) and denote the \( m^{th} \) result as \( p_{i,m}^* \).

5. If there are \( I \) iterations, return \( \frac{1}{M} \sum_m \frac{1}{I} \sum_i p_{i,m}^* \) for the posterior predictive probability.
B.4 MCMC Sampling Proofs

B.4.1 HMC

B.4.1.1 SvM-c

Lemma B.4.1. The derivatives for $\arctan^*(z_1, z_2)$ are

$$
\frac{d}{dz_1} \arctan^*(z_1, z_2) = \frac{-z_2}{z_1^2 + z_2^2}, \\
\frac{d}{dz_2} \arctan^*(z_1, z_2) = \frac{z_1}{z_1^2 + z_2^2}.
$$

Proof. By the chain rule, the derivative for $\arctan^*(z_1, z_2)$ first involves taking the derivative of $\arctan\left(\frac{z_2}{z_1}\right)$. The rest follows by taking the appropriate derivative of $\frac{z_2}{z_1}$ and suitable rearrangement of the gradient.

Lemma B.4.2. For the model specified in (B.3) and using the parametrization given in (B.15), the update for the momentum vector are respectively

$$
- \frac{\partial U}{\partial q_{m\ell}}(q(t)) = \left(\rho_{\ell} \cos(y_{\ell}) + r_{\ell} \left(\Sigma^{-1}(r \cos(m) - \mu_1)\right)_{\ell}\right) \sin(m_{\ell}) - \\
\left(\rho_{\ell} \sin(y_{\ell}) + r_{\ell} \left(\Sigma^{-1}(r \sin(m) - \mu_2)\right)_{\ell}\right) \cos(m_{\ell}),
$$

(B.36)

$$
- \frac{\partial U}{\partial q_{r\ell}}(q(t)) = \frac{1}{r_{\ell}} \left(\Sigma^{-1}(r \cos(m) - \mu_1)\right)_{\ell} \cos(m_{\ell}) - \\
\left(\Sigma^{-1}(r \sin(m) - \mu_1)\right)_{\ell} \sin(m_{\ell}).
$$

(B.37)

Proof. Suppressing all other parts that do not depend on $m_{\ell}$ and $r_{\ell}$, we have that

$$
\log(p(m_{\ell}, r_{\ell} | y)) = \rho_{\ell} \cos(y_{\ell} - m_{\ell}) + \sum_{\ell=1}^{N} \log(r_{\ell}) - \frac{1}{2} (r \cos(m) - \mu_1)^T \Sigma^{-1}(r \cos(m) - \mu_1) - \\
\frac{1}{2} (r \sin(m) - \mu_2)^T \Sigma^{-1}(r \sin(m) - \mu_2) + C.
$$

Without loss of generality, let $v = r \cos(m) - \mu_1$. Note that $\frac{dv_{\ell}}{dm_{\ell}} = -r_{\ell} \sin(m_{\ell})$ and $\frac{dv_{\ell}}{dr_{\ell}} = \cos(m_{\ell})$ for $\ell' = \ell$ and 0 otherwise. Hence,

$$
\frac{d}{dm_{\ell}} \frac{1}{2} (r \cos(m) - \mu_1)^T \Sigma^{-1}(r \cos(m) - \mu_1) = \left(\Sigma^{-1}(r \cos(m) - \mu_1)\right)_{\ell} - r_{\ell} \sin(m_{\ell})
$$
and

\[
\frac{d}{dr_\ell} \left( r \cos(m) - \mu_1 \right)^T \Sigma^{-1} (r \cos(m) - \mu_1) = \left( \Sigma^{-1} (r \cos(m) - \mu_1) \right)_\ell \cos(m_\ell).
\]

A similar result holds for \( \frac{1}{2} (r \sin(m) - \mu_2)^T \Sigma^{-1} (r \sin(m) - \mu_2) \).

If we then take the derivative of \( \log(r_\ell) \), we get the quantity in B.37. On the other hand, for the quantity in B.36,

\[
\frac{d}{dm_\ell} \rho_\ell \cos(y_\ell - m_\ell) = -\rho_\ell \sin(y_\ell - m_\ell)
\]

\[
= -\rho_\ell (\sin(y_\ell) \cos(m_\ell) - \sin(m_\ell) \cos(y_\ell)).
\]

Some rearrangement then gives us the quantity in B.36.

\[\square\]

Lemma B.4.3. For the model specified in (B.3) and using the parametrization given in (B.18), the update for the momentum vectors are respectively

\[
- \frac{\partial U}{\partial q_{\tilde{m}e}} (q(t)) = 
\sum_{\ell=1}^N - \rho_{\ell} \sin(y_{\ell} - m_{\ell}) \left( \frac{-z_{1,\ell}'}{z_{1,\ell}' + z_{2,\ell}' L_{\ell,\ell} \tilde{r}_{\ell} \sin(\tilde{m}_{\ell})} + \frac{z_{1,\ell}'}{z_{1,\ell}' + z_{2,\ell}' L_{\ell,\ell} \tilde{r}_{\ell} \cos(\tilde{m}_{\ell})} \right), \quad \text{(B.38)}
\]

\[
- \frac{\partial U}{\partial q_{r_\ell}} (q(t)) = 
\sum_{\ell=1}^N - \rho_{\ell} \sin(y_{\ell} - m_{\ell}) \left( \frac{-z_{2,\ell}'}{z_{1,\ell}' + z_{2,\ell}' L_{\ell,\ell} \cos(\tilde{m}_{\ell})} + \frac{z_{1,\ell}'}{z_{1,\ell}' + z_{2,\ell}' L_{\ell,\ell} \sin(\tilde{m}_{\ell})} \right) + \frac{1}{r_{\ell}} - \tilde{r}_{\ell}. \quad \text{(B.39)}
\]

Proof. Using the chain rule, we have that

\[
\frac{d}{dm_\ell} \rho_{\ell} \cos(y_{\ell} - m_{\ell}) = \left( \frac{d}{dm_\ell} \rho_{\ell} \cos(y_{\ell} - m_{\ell}) \right) \left( \frac{dm_\ell}{dz_{1,\ell}'} \frac{dz_{1,\ell}'}{dz_{1,\ell}} \frac{d\tilde{m}_e}{dm_\ell} + \frac{dm_\ell}{dz_{2,\ell}'} \frac{dz_{2,\ell}'}{dz_{2,\ell}} \frac{d\tilde{m}_e}{dm_\ell} \right)
\]

and

\[
\frac{d}{dr_\ell} \rho_{\ell} \cos(y_{\ell} - m_{\ell}) = \left( \frac{d}{dm_\ell} \rho_{\ell} \cos(y_{\ell} - m_{\ell}) \right) \left( \frac{dm_\ell}{dz_{1,\ell}'} \frac{dz_{1,\ell}'}{dz_{1,\ell}} \frac{d\tilde{r}_e}{dr_\ell} + \frac{dm_\ell}{dz_{2,\ell}'} \frac{dz_{2,\ell}'}{dz_{2,\ell}} \frac{d\tilde{r}_e}{dr_\ell} \right).
\]

Again, \( \frac{dz_{1,\ell}'}{dm_\ell} = -r_\ell \sin(m_{\ell}) \) and \( \frac{dz_{2,\ell}'}{dm_\ell} = r_\ell \cos(m_{\ell}) \) if \( \ell' = \ell \) and 0 otherwise. A similar result holds for \( \frac{dz_{1,\ell}'}{dr_\ell} \) and \( \frac{dz_{2,\ell}'}{dr_\ell} \). Then, by definition, \( \frac{dz_{1,\ell}'}{dz_{2,\ell}'} = \frac{dz_{2,\ell}'}{dz_{1,\ell}'} = L_{j,i} \). Finally, Lemma B.4.1 gives us \( \frac{dm_\ell}{dz_{1,\ell}'} \) and \( \frac{dm_\ell}{dz_{2,\ell}'} \).
Because the distribution on $m_\ell$ is uniform, We get the quantity in B.38 by summing over all observations. For $r_\ell$, we also have to add in the derivative of $\log(r_\ell) - \frac{1}{2} r_\ell^2$. 

### B.4.1.2 SvM-p

**Lemma B.4.4.** For the model specified in (3.16), the update for the momentum vector corresponding to $z_\ell$ in the centered parametrization is

$$
\lambda_{k,\ell} \frac{f_V(y_\ell \mid m_k, \rho_k)}{p(y_\ell \mid z, m, \rho)} - \sum_{k'=1}^{K} \lambda_{k',\ell} \frac{f_V(y_\ell \mid m_{k'}, \rho_{k'})}{p(y_\ell \mid z, m, \rho)} - (\Sigma^{-1} z)_\ell.
$$

**Proof.** We have that

$$
\frac{d}{dz_{k,\ell}} \log(p(z, m, \rho \mid y)) \approx \frac{d}{dz_{k,\ell}} \log(p(y_\ell \mid z, m, \rho)) + \log(GP(z \mid 0, \Sigma))
$$

$$
= \sum_{k'=1}^{K} \frac{d \lambda_{k',\ell}}{dz_{k,\ell}} \frac{d}{d \lambda_{k',\ell}} \log(p(y_\ell \mid z, m, \rho)) + \frac{d}{dz_{\ell}} \log(GP(z \mid 0, \Sigma))
$$

$$
= \frac{\sum_{k'=1}^{K} f_V(y_\ell \mid m_{k'}, \rho_{k'}) \frac{d \lambda_{k',\ell}}{dz_{k,\ell}}}{p(y_\ell \mid z, m, \rho)} - (\Sigma^{-1} z)_\ell.
$$

If $k' = k$, then

$$
\frac{d \lambda_{k,\ell}}{dz_{k,\ell}} = \frac{d}{dz_{k,\ell}} \frac{\exp(z_{k,\ell})}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})}
$$

$$
= \frac{\exp(z_{k,\ell})}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})} = \left( \frac{\exp(z_{k,\ell})}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})} \right)^2
$$

$$
= \lambda_{k,\ell}(1 - \lambda_{k,\ell}).
$$

If $k' \neq k, k' \neq K$ then

$$
\frac{d \lambda_{k',\ell}}{dz_{k',\ell}} = \frac{d}{dz_{k',\ell}} \frac{\exp(z_{k',\ell})}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})}
$$

$$
= -\frac{\exp(z_{k,\ell}) \exp(z_{k',\ell})}{\left(1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})\right)^2}
$$

$$
= -\lambda_{k,\ell} \lambda_{k',\ell}.
$$
If \( k' = K \), then

\[
\frac{d\lambda_{k',\ell}}{dz_{k',\ell}} = \frac{d}{dz_{k',\ell}} \frac{1}{1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell})} = -\exp(z_{k',\ell}) \left( 1 + \sum_{k'=1}^{K-1} \exp(z_{k',\ell}) \right)^2 = -\lambda_{k,\ell} \lambda_{K,\ell}.
\]

**Lemma B.4.5.** For the model specified in (3.17), the update for the momentum vector corresponding to \( z_{\ell} \) in the centered parametrization is

\[
\lambda_{1,\ell} (1 - \lambda_{1,\ell}) \frac{f_V (y_{\ell} \mid m_1, \rho_1) - f_V (y_{\ell} \mid m_2, \rho_2)}{p(y_{\ell} \mid z_{\ell}, m_1, m_2, \rho_1, \rho_2)} - (\Sigma^{-1}z)_{\ell}.
\]

**Proof.** We have that

\[
\frac{d}{dz_{\ell}} \log(p(z, m_1, m_2, \rho_1, \rho_2 \mid y)) = \frac{d}{dz_{\ell}} \log(p(y_{\ell} \mid z, m_1, m_2, \rho_1, \rho_2)) + \log(\text{GP}(z \mid 0, \Sigma))
\]

\[
= \frac{d\lambda_{\ell}}{dz_{\ell}} \frac{d}{d\lambda_{\ell}} \log(p(y_{\ell} \mid z, m_1, m_2, \rho_1, \rho_2)) + \frac{d}{dz_{\ell}} \log(\text{GP}(z \mid 0, \Sigma))
\]

\[
= \frac{d\lambda_{\ell}}{dz_{\ell}} \frac{f_V (y_{\ell} \mid m_1, \rho_1) - f_V (y_{\ell} \mid m_2, \rho_2)}{p(y_{\ell} \mid z_{\ell}, m_1, m_2, \rho_1, \rho_2)} - (\Sigma^{-1}z)_{\ell}.
\]

As \( \lambda_{\ell} = \psi^{-1}(z_{\ell}) \),

\[
\frac{d\lambda_{\ell}}{dz_{\ell}} = \frac{d}{dz_{\ell}} \frac{1}{1 + \exp(-z_{\ell})}
\]

\[
= \frac{-1}{(1 + \exp(-z_{\ell}))^2} - \exp(-z_{\ell})
\]

\[
= \frac{\exp(-z_{\ell})}{1 + \exp(-z_{\ell})} \frac{1}{1 + \exp(-z_{\ell})}
\]

\[
= \left( 1 - \frac{1}{1 + \exp(-z_{\ell})} \right) \frac{1}{1 + \exp(-z_{\ell})}
\]

\[
= (1 - \psi^{-1}(z_{\ell})) \psi^{-1}(z_{\ell}).
\]

Putting these terms together gives us the quantity in the lemma. \( \square \)

**Lemma B.4.6.** For the model specified in 3.16, the update for the momentum vector corresponding
As before, we need $d\psi$ to $\tilde{z}_t$ in the noncentered parametrization is

$$
\sum_{\ell'} L'_{\ell',\ell} \lambda_{k,\ell'} \frac{f_{\ell'}(y_{\ell'} \mid m_k, \rho_k)}{p(y_{\ell'} \mid z_{\ell'}, m, \rho)} - \sum_{k'=1}^{K} \lambda_{k',\ell} \frac{f_{\ell'}(y_{\ell'} \mid m_{k'}, \rho_{k'})}{p(y_{\ell'} \mid z_{\ell'}, m, \rho)} - \tilde{z}_t.
$$

Proof. We have that

$$
\frac{d}{d\tilde{z}_t} \log(P(z, m, \rho \mid y))
\propto \frac{d}{d\tilde{z}_t} \log(P(y_{\ell'} \mid z, m_1, m_2, \rho_1, \rho_2)) + \log(N(\tilde{z}_t \mid 0, 1))
= \sum_{\ell'} \sum_{k'=1}^{K} \frac{d\tilde{z}_t}{d\tilde{z}_t} \frac{d\lambda_{k',\ell}}{d\lambda_{k',\ell}} \frac{d}{d\lambda_{k',\ell}} \log(P(y_{\ell'} \mid z, m_1, m_2, \rho_1, \rho_2)) - \tilde{z}_t
= \sum_{\ell'} \sum_{k'=1}^{K} \frac{d\tilde{z}_t}{d\tilde{z}_t} \frac{d\lambda_{k',\ell}}{d\lambda_{k',\ell}} \frac{d}{d\lambda_{k',\ell}} f_{\ell'}(y_{\ell'} \mid m'_{k'}, \rho'_{k'}) - \tilde{z}_t.
$$

As before, we need $\frac{d\psi_{\ell'}}{d\tilde{z}_t}$ and $\frac{d\tilde{z}_t}{d\tilde{z}_t}$ for the chain rule. We get $\frac{d\lambda_{\ell'}}{d\tilde{z}_t}$ from before. Then, as $z_{\ell'} = L_{\ell, \tilde{z}}$, $\frac{dz_{\ell'}}{dz_{\ell}} = L'_{\ell',\ell}$. Putting these terms together gives us the quantity in the lemma.

Lemma B.4.7. For the model specified in 3.17, the update for the momentum vector corresponding to $\tilde{z}_t$ in the noncentered parametrization is

$$
\sum_{\ell'} (1 - \lambda_{1,\ell'}) \frac{f_{\ell'}(y_{\ell'} \mid m_k, \rho_k)}{p(y_{\ell'} \mid z_{\ell'}, m, \rho)} - \tilde{z}_t.
$$

Proof. We have that

$$
\frac{d}{d\tilde{z}_t} \log(P(z, m_1, m_2, \rho_1, \rho_2 \mid y))
\propto \frac{d}{d\tilde{z}_t} \log(P(y_{\ell'} \mid z, m_1, m_2, \rho_1, \rho_2)) + \log(N(\tilde{z}_t \mid 0, 1))
= \sum_{\ell'} \frac{d\tilde{z}_t}{d\tilde{z}_t} \frac{d\lambda_{1,\ell'}}{d\lambda_{1,\ell'}} \frac{d}{d\lambda_{1,\ell'}} \log(P(y_{\ell'} \mid z, m_1, m_2, \rho_1, \rho_2)) - \tilde{z}_t
= \sum_{\ell'} \frac{d\tilde{z}_t}{d\tilde{z}_t} \frac{d\lambda_{1,\ell'}}{d\lambda_{1,\ell'}} \frac{d}{d\lambda_{1,\ell'}} f_{\ell'}(y_{\ell'} \mid m_k, \rho_k) - \tilde{z}_t.
$$

As before, we need $\frac{d\psi_{\ell'}}{d\tilde{z}_t}$ and $\frac{d\tilde{z}_t}{d\tilde{z}_t}$ for the chain rule. We get $\frac{d\lambda_{\ell'}}{d\tilde{z}_t}$ from before. Then, as $z_{\ell'} = L_{\ell, \tilde{z}}$, $\frac{dz_{\ell'}}{dz_{\ell}} = L'_{\ell',\ell}$. Putting these terms together gives us the quantity in the lemma.
B.4.2 Expectation maximization

B.4.2.1 SvM-c

**Lemma B.4.8.** The $\nu_k$ that maximizes the conditional log posterior in (B.30) is

$$
\sum_{\ell} \frac{r_{k,\ell} \varphi_{k,\ell}}{\varsigma^2} \sum_{\ell} \frac{r_{k,\ell}}{\varsigma^2} + \frac{1}{\tau^2}.
$$

**(B.40)**

**Proof.** Gathering the terms that depend on $\nu_k$ and taking the derivative, we have that

$$
\frac{d}{d\nu_k} \left( \sum_{\ell} - \frac{r_{k,\ell}}{2 \varsigma^2} (\varphi_k - \nu_k)^2 - \frac{1}{2 \tau^2} \nu_k^2 \right) = 0
$$

$$
\sum_{\ell} \frac{r_{k,\ell}}{\varsigma^2} (\varphi_k - \nu_k) - \frac{1}{\tau^2} \nu_k = 0
$$

$$
\nu_k \left( \sum_{\ell} \frac{r_{k,\ell}}{\varsigma^2} - \frac{1}{\tau^2} \right) = \sum_{\ell} \frac{r_{k,\ell}}{\varsigma^2} \varphi_k
$$

$$
\nu_k = \frac{\sum_{\ell} \frac{r_{k,\ell} \varphi_{k,\ell}}{\varsigma^2}}{\sum_{\ell} \frac{r_{k,\ell}}{\varsigma^2} + \frac{1}{\tau^2}}.
$$

**Lemma B.4.9.** The gradient of $\varphi_{k,\ell}$ with respect to the conditional log posterior in (B.30) is

$$
\rho_{k,\ell} \left( \cos(y_{\ell} - m_{k,\ell}) - \frac{I_1(\rho_{k,\ell})}{I_0(\rho_{k,\ell})} \right) - \frac{1}{\varsigma^2} (\varphi_{k,\ell} - \nu_k).
$$

**Proof.** Gathering the terms that depend on $\varphi_{k,\ell}$, we have that

$$
r_{k,\ell} \rho_{k,\ell} \cos(y_{\ell} - m_{k,\ell}) - \log(I_0(\rho_{k,\ell})) - \frac{1}{2 \varsigma^2} (\varphi_{k,\ell} - \nu_k)^2.
$$

Then, because

$$
\frac{d}{d\varphi_i} \rho_{k,\ell} = \frac{d}{d\varphi_i} \exp(\varphi_{k,\ell}) = \exp(\varphi_{k,\ell}) = \rho_{k,\ell}
$$

and

$$
\frac{d}{d\rho_{k,\ell}} \log(I_0(\rho_{k,\ell})) = \frac{I_{-1}(\rho_{k,\ell})}{I_0(\rho_{k,\ell})},
$$

according to Wolfram Research [73], we get the result in the lemma from the chain rule.

**Lemma B.4.10.** The gradient of $\tilde{z}_{k,1,\ell}$ and $\tilde{z}_{k,2,\ell}$ from the conditional log posterior given in (B.30)

169
\[ \frac{\partial}{\partial \tilde{z}_{k,1,\ell}} \mathbb{E} \left( \log \left( \tilde{p}(\mathbf{z}, \Theta \mid \mathbf{y}, \mathbf{x}) \right) \right) = \sum_{e'} \tilde{L}_{e',r_{k,e'}} \frac{-z_{k,2,e'}}{z_{k,1,e'}^2 + z_{k,2,e'}^2} \rho_{k,e'} \sin(y_{e'} - m_{k,e'}) - \tilde{z}_{k,\ell}. \]

\[ \frac{\partial}{\partial \tilde{z}_{k,2,\ell}} \mathbb{E} \left( \log \left( \tilde{p}(\mathbf{z}, \Theta \mid \mathbf{y}, \mathbf{x}) \right) \right) = \sum_{e'} \tilde{L}_{e',r_{k,e'}} \frac{z_{k,1,e'}}{z_{k,1,e'}^2 + z_{k,2,e'}^2} \rho_{k,e'} \sin(y_{e'} - m_{k,e'}) - \tilde{z}_{k,\ell}. \]

**Proof.** Without loss of generality, gathering the terms that depend on \( \tilde{z}_{k,1,\ell} \), we have that the derivative is

\[ \sum_{e'} (r_{k,e'} \rho_{k,e'} \sin(y_{e'} - m_{k,e'})) \frac{dm_{k,e'}}{dz_{k,1,e'}} \frac{dz_{k,1,e'}}{d\tilde{z}_{k,1,\ell}} - \tilde{z}_{k,1,\ell}. \]

We need \( \frac{dm_{k,e'}}{dz_{k,1,e'}} \) and \( \frac{dz_{k,1,e'}}{d\tilde{z}_{k,1,\ell}} \) for the chain rule. As before, because \( z_{k,1,e'} = L_{e'} \cdot \tilde{z}_{k,1} \), \( \frac{dz_{k,1,e'}}{d\tilde{z}_{k,1,\ell}} = L_{e',\ell} \).

Then, because \( m_{k,e'} = \arctan^*(z_{k,1,e'}, z_{k,2,e'}) \) and if \( C \) is a constant in \( \mathbb{R} \),

\[ \frac{dm_{k,e'}}{dz_{k,1,e'}} = \frac{d}{dz_{k,1,e'}} \arctan(z_{k,1,e'}; z_{k,2,e'}) + C \]

\[ = \frac{1}{1 + \left( \frac{z_{k,2,e'}}{z_{k,1,e'}} \right)^2} \left( -\frac{z_{k,2,e'}}{z_{k,1,e'}} \right) \]

\[ = -\frac{z_{k,2,e'}}{z_{k,1,e'}^2 + z_{k,2,e'}^2}. \]

Meanwhile, following a similar calculation, we get that

\[ \frac{dm_{k,e'}}{dz_{k,2,e'}} = \frac{z_{k,1,e'}}{z_{k,1,e'}^2 + z_{k,2,e'}^2}. \]

Putting these terms together gives us the quantity in the lemma. \( \square \)

### B.4.2.2 SvM-p

**Lemma B.4.11.** The gradient of \( \tilde{z}_{k,\ell} \) from the conditional log posterior in (B.32) is

\[ \sum_{e'} (r_{k,e'} - \lambda_{k,e'}) L_{e',\ell} - \tilde{z}_{k,\ell}. \]
Proof. Gathering terms that depend on \( e_z k,\ell \) and using work from earlier, we have that

\[
\frac{\partial}{\partial e_z k,\ell} \mathbb{E} (\log(p(\zeta, \Theta | y, x))) = \sum e e \sum_{e'} \frac{\partial z_{k,e'}}{\partial e_z k,\ell} \frac{\partial \lambda z_{k,e'}}{\partial \lambda z_{k,e'}} \frac{\partial}{\partial r_{k',e}} \frac{r_{k',e}}{\lambda_{k',e'}} \log(\lambda_{k',e'}) = \frac{1}{2} e_z k,\ell.
\]

\[
\sum e e \sum_{e'} \frac{\partial z_{k,e'}}{\partial e_z k,\ell} \frac{\partial \lambda z_{k,e'}}{\partial \lambda z_{k,e'}} \frac{\partial}{\partial r_{k',e}} \frac{r_{k',e}}{\lambda_{k',e'}} \log(\lambda_{k',e'}) = \left(1 - r_{1,k} \right) \log(1 - \lambda_{1,\ell}) - \left(1 - r_{1,k} \right) \lambda_{1,\ell}.
\]

Lemma B.4.12. The gradient of \( \tilde{z}_{k,\ell} \) from the conditional log posterior in (B.32) is

\[
\sum e e \left(1 - r_{1,k} \right) \log(1 - \lambda_{1,\ell}) - \left(1 - r_{1,k} \right) \lambda_{1,\ell}.
\]

Proof. Gathering the terms that depend on \( \tilde{z}_{k,\ell} \), we have that

\[
\sum e e r_{1,k} \log(\lambda_{1,\ell}) + \left(1 - r_{1,k} \right) \log(1 - \lambda_{1,\ell}) = \left(1 - r_{1,k} \right) \lambda_{1,\ell}.
\]

As before, we need \( \frac{d\lambda_{1,e'}}{dz_{\ell}} \) and \( \frac{dz_{\ell}}{dz_{k,e'}} \) for the chain rule. The latter is still the same, i.e. \( L_{e',\ell} \). Then, because \( \lambda_{1,e'} = \psi^{-1}(z_{e'}) \), we have from earlier that

\[
\frac{d\lambda_{1,e'}}{dz_{\ell}} = (1 - \lambda_{1,\ell})(\lambda_{1,e'}).
\]

As a result, we get that

\[
\frac{d}{dz_{\ell}} r_{1,k} \log(\lambda_{1,\ell}) + \left(1 - r_{1,k} \right) \log(1 - \lambda_{1,\ell}) = r_{1,k}(1 - \lambda_{1,\ell}) - (1 - r_{1,k}) \lambda_{1,\ell},
\]

which simplifies to the quantity in the lemma. \( \square \)
B.5 Additional simulation and results

B.5.1 Real data duplicates and posterior plots

Figure B.1f shows us that the data violates a key assumption. There are locations with observed angles that are seen multiple times in the data set. Due to the use of a continuous distribution for the observations, i.e. the von Mises distribution, the probability of this happening should be zero. It is more problematic for the \( S_vM - c \) model if we use the hierarchical prior for the concentration parameter because repeated observations might induce large values, which will pull up all values for the concentration parameter.

Fortunately, it is straightforward to identify these duplicates. Figure B.1 shows how observations are duplicated if we break it down by number, location, and/or the value of the observed angle. Despite there being some variation in the exact value, most duplicate observations are only observed twice. Further, the proportion of duplicated observations decreases over time because it appears to oscillate around 0.12 before 2002-2003 and then fluctuates around 0.07 afterwards. Interestingly enough, this decrease in the proportion also roughly corresponds to a spike in the maximum number of duplicates. This suggests a concentration of duplicates at fewer locations, which is observed in Figure B.1b from 2004 onward. However, the maximum number of duplicate combinations by year and the duplicate by locations appear to be correlated and exhibit trends that are independent of the number of observations. The maximum number of duplicates peaks at 36 in 1994-1995 before coming back down. Simultaneously, the locations with duplicates goes from a diffuse spread across the simplex to a concentrated spread on the lower half of the boundary in which the proportion of the third income category is close to zero. It then returns to a more diffuse set of locations by 2003-2004. Finally, while there is a lot of noise in the duplicated observed direction value across the years, there does appear to be a peak around \( \frac{\pi}{2} \) and \( \frac{3\pi}{2} \). This is seen in the plot for observed directions and their locations weighted by the number of appearances for 2008-2009.

B.5.2 Real data results

B.5.3 Additional simulation plots
<table>
<thead>
<tr>
<th>Year (Model)</th>
<th>$m$</th>
<th>$\rho$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2001</td>
<td>$\pi_1 = 5.97$ (5.39, 0.25)</td>
<td>$\pi_2 = 5.51$ (5.30, 0.70)</td>
<td>$\rho_1 = 1.10$ (0.90, 1.31)</td>
</tr>
<tr>
<td>2001-2002</td>
<td>$\pi_1 = 1.34$ (0.95, 1.62)</td>
<td>$\pi_2 = 2.77$ (1.82, 3.72)</td>
<td>$\rho_1 = 0.97$ (0.50, 1.72)</td>
</tr>
<tr>
<td>2002-2003</td>
<td>$\pi_1 = 5.28$ (5.11, 5.45)</td>
<td>$\pi_2 = 2.78$ (2.30, 3.35)</td>
<td>$\rho_1 = 0.79$ (0.73, 0.84)</td>
</tr>
<tr>
<td>2003-2004</td>
<td>$\pi_1 = 1.40$ (1.03, 1.75)</td>
<td>$\pi_2 = 3.00$ (2.14, 3.96)</td>
<td>$\rho_1 = 2.57$ (2.43, 3.42)</td>
</tr>
<tr>
<td>2004-2005</td>
<td>$\pi_1 = 5.15$ (4.81, 5.12)</td>
<td>$\pi_2 = 2.87$ (2.43, 3.42)</td>
<td>$\rho_1 = 0.05$ (0.29, 0.61)</td>
</tr>
<tr>
<td>2005-2006</td>
<td>$\pi_1 = 1.56$ (1.13, 1.83)</td>
<td>$\pi_2 = 5.38$ (5.16, 5.64)</td>
<td>$\rho_1 = 1.36$ (0.96, 1.71)</td>
</tr>
<tr>
<td>2006-2007</td>
<td>$\pi_1 = 5.15$ (4.48, 5.85)</td>
<td>$\pi_2 = 1.87$ (1.35, 2.43)</td>
<td>$\rho_1 = 2.37$ (1.90, 2.95)</td>
</tr>
<tr>
<td>2007-2008</td>
<td>$\pi_1 = 2.29$ (1.48, 3.32)</td>
<td>$\pi_2 = 4.64$ (4.16, 5.06)</td>
<td>$\rho_1 = 1.65$ (1.22, 2.25)</td>
</tr>
<tr>
<td>2008-2009</td>
<td>$\pi_1 = 2.25$ (1.69, 2.77)</td>
<td>$\pi_2 = 4.70$ (4.33, 5.05)</td>
<td>$\rho_1 = 1.70$ (1.24, 2.42)</td>
</tr>
<tr>
<td>2009-2010</td>
<td>$\pi_1 = 2.00$ (1.73, 2.29)</td>
<td>$\pi_2 = 4.92$ (4.27, 5.58)</td>
<td>$\rho_1 = 2.05$ (1.49, 2.76)</td>
</tr>
</tbody>
</table>

Table B.1: Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the income proportions in Los Angeles County from 1990 to 2010. Parameters with bars over them were averaged across all locations.
Figure B.1: Plots in the top row show the duplicate observations broken down by different categories across different years with further context provided by plots in the bottom row.

Figure B.2: Plots showing the posterior predictive log probability defined in (3.34) and the von Mises distributions’ parameters and the mixing probability of the first component selected according to that probability and the models’ mixing for each year. For $SvM$ and $SvM-c$, the mean and concentration parameters’ mean and credible interval are averaged over all locations. On the other hand, the mean and credible interval for the probability of the first component is averaged across all location for $SvM-p$. In addition, the second component’s values are shown with an opacity of 0.5.
Figure B.3: Plots showing the observed random direction not withheld and the fitted mean surface of the model selected by the posterior predictive log probability in (3.34). The front axis represents the proportion in the first income category and the side axis represents the proportion in the second income category. The up-down axis represents the direction. The start of the arrow indicates a value of zero whereas the end indicates a value of $2\pi$ for angles and 1 for proportions.
Table B.2: Time it takes for our models to run four chains for different models in seconds according to the sampling scheme outlined in Section 3.5.1 on four cores of a cluster using 2x 3.0 GHz Intel Xeon Gold 6154 as its processor.

<table>
<thead>
<tr>
<th>Model</th>
<th>( iV )</th>
<th>( iVM )</th>
<th>( \text{SvM mean } \pi )</th>
<th>( \text{SvM-c} )</th>
<th>( \text{SvM-p} )</th>
<th>( \text{SvM mean 0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{SvM} ) cf. (B.3)</td>
<td>4528</td>
<td>3039</td>
<td>5209</td>
<td>2663</td>
<td>3376</td>
<td>5488</td>
</tr>
<tr>
<td>( \text{SvM-c} ) cf. (3.15)</td>
<td>23738</td>
<td>16442</td>
<td>29486</td>
<td>22624</td>
<td>16546</td>
<td>30571</td>
</tr>
<tr>
<td>( \text{SvM-p} ) cf. (3.16)</td>
<td>813</td>
<td>1188</td>
<td>816</td>
<td>556</td>
<td>866</td>
<td>535</td>
</tr>
</tbody>
</table>
Figure B.4: Plots of the mean component fitted by different models for observations simulated according to the $SvM$ model in (B.3). The fitted mean components are shown in colored asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations.

<table>
<thead>
<tr>
<th></th>
<th>$m$</th>
<th>$\rho$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>$\bar{m} = 3.23 (1.78, 4.86)$</td>
<td>$\bar{\rho} = 3.00 (2.54, 3.50)$</td>
<td>$-$</td>
</tr>
<tr>
<td>$SvM$ cf. (B.3)</td>
<td>$\bar{m} = 3.28 (2.89, 3.68)$</td>
<td>$\bar{\rho} = 3.14 (2.70, 3.63)$</td>
<td>$-$</td>
</tr>
<tr>
<td>$SvM$-c cf. (3.15)</td>
<td>$\bar{m}_1 = 1.58 (0.39, 2.78)$</td>
<td>$\bar{\rho}_1 = 0.37 (0.00, 89.98)$</td>
<td>$\lambda_1 = 0.02 (9.07e-4, 0.04)$</td>
</tr>
<tr>
<td></td>
<td>$\bar{m}_2 = 3.32 (2.90, 3.76)$</td>
<td>$\bar{\rho}_2 = 3.45 (2.93, 4.09)$</td>
<td>$\lambda_2 = 0.98 (0.96, 0.9991)$</td>
</tr>
<tr>
<td>$SvM$-p cf. (3.16)</td>
<td>$m_1 = 2.94 (2.78, 3.10)$</td>
<td>$\rho_1 = 2.90 (2.15, 3.78)$</td>
<td>$\bar{\lambda}_1 = 0.51 (0.31, 0.72)$</td>
</tr>
<tr>
<td></td>
<td>$m_2 = 3.81 (3.60, 3.99)$</td>
<td>$\rho_2 = 1.65 (1.15, 2.30)$</td>
<td>$\bar{\lambda}_2 = 0.49 (0.28, 0.69)$</td>
</tr>
</tbody>
</table>

Table B.3: Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.4. Parameters with bars over them were averaged across all locations.
Figure B.5: Plots of the mean component fitted by different models for observations simulated according to the $SvM$ model in (B.3). The fitted mean components are shown in colored asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations.

<table>
<thead>
<tr>
<th>Model</th>
<th>$m$</th>
<th>$\rho$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simulation</strong></td>
<td>$\bar{m} = 0.68$ (4.17, 2.38)</td>
<td>$\bar{\rho} = 2.99$ (2.51, 3.45)</td>
<td>---</td>
</tr>
<tr>
<td>$SvM$ cf. (B.3)</td>
<td>$\bar{m} = 0.64$ (6.27, 1.27)</td>
<td>$\bar{\rho} = 2.65$ (2.24, 3.12)</td>
<td>---</td>
</tr>
<tr>
<td>$SvM$-c cf. (3.15)</td>
<td>$m_1 = 0.67$ (0.10, 1.24)</td>
<td>$\bar{\rho}_1 = 2.66$ (2.24, 3.19)</td>
<td>$\lambda_1 = 0.99$ (0.95, 0.9996)</td>
</tr>
<tr>
<td></td>
<td>$m_2 = 4.72$ (3.52, 5.91)</td>
<td>$p_2 = 0.12$ (3.40e-5, 225.56)</td>
<td>$\lambda_2 = 0.01$ (4.51e-4, 0.05)</td>
</tr>
<tr>
<td>$SvM$-p cf. (3.16)</td>
<td>$m_1 = 1.48$ (1.34, 1.61)</td>
<td>$\bar{\rho}_1 = 2.18$ (1.77, 2.64)</td>
<td>$\bar{\lambda}_1 = 0.53$ (0.31, 0.75)</td>
</tr>
<tr>
<td></td>
<td>$m_2 = 5.64$ (5.46, 5.83)</td>
<td>$\rho_2 = 1.60$ (1.26, 1.98)</td>
<td>$\bar{\lambda}_2 = 0.47$ (0.25, 0.69)</td>
</tr>
</tbody>
</table>

Table B.4: Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.4. Parameters with bars over them were averaged across all locations.
Figure B.6: Plots of the mean component fitted by different models for observations simulated according to the $SVM-p$ model (B.3) in the bottom row. The fitted mean components are shown in red asterisks with the simulated observations as circles and the simulated means as grey rhombuses for different scenarios. The x-y axis are the observations’ location in the first two dimensions of a three dimensional simplex; the z-axis represent the observations.

| Simulation | $m_1$ = 1.57 | $m_2$ = 4.71 | $\lambda_1$ = 0.51 (0.16, 0.91) | $\lambda_2$ = 0.49 (0.09, 0.84) |
| SvM cf. (B.3) | $m$ = 3.22 (1.87, 4.56) | $\rho_1$ = 5 | $\rho_2$ = 10 |
| SvM-c cf. (3.15) | $\overline{m}_1$ = 1.63 (1.23, 2.02) | $\overline{m}_2$ = 4.72 (4.34, 5.08) | $\lambda_1$ = 0.53 (0.48, 0.57) | $\lambda_2$ = 0.47 (0.43, 0.52) |
| SvM-p cf. (3.16) | $m_1$ = 1.62 (1.57, 1.68) | $m_2$ = 4.71 (4.67, 4.75) | $\rho_1$ = 5.36 (4.55, 6.25) | $\rho_2$ = 9.85 (8.19, 11.67) | $\lambda_1$ = 0.52 (0.30, 0.73) | $\lambda_2$ = 0.48 (0.27, 0.70) |

Table B.5: Circular mean posterior values and 95% credible intervals for the parameters in the Von Mises distribution and mixing probability are shown for the models fitted to the simulation data in Figure B.6. Parameters with bars over them were averaged across all locations.
APPENDIX C

Tree-Based Mixture of Simplices Appendix

C.1 Proofs

C.1.1 Metrics

**Lemma C.1.1.** If $T$ and $T'$ are two trees such that the conditions hold in Definition 4.3.2 and all vertices are distinct, then $d_P(T, T')$ is a valid metric.

**Proof.** Note that because of assumptions (A2) and (A3), the path permutation is a bijective mapping. Further, this metric is clearly non-negative because the sum of Hausdorff metrics must be non-negative. We now prove the three properties of the metric.

$d_P(T, T') = 0 \iff T = T'$: Because the Hausdorff metric is a proper metric and the path permutation is bijective, $d_P(T, T') = 0 \iff d_H(G_{\varphi^i}, G'_{\sigma_P(\varphi^i)}) = 0$ for all $i = 1, 2, \ldots, I$. This does not necessarily imply that $T = T'$ due to the trees imposing an order on the topics in the convex hull and the Hausdorff metric ignoring this order. However, we can use contradiction to show that this is implied. Suppose for some $i$, $d_H(G_{\varphi^i}, G'_{\sigma_P(\varphi^i)}) = 0$, but $\varphi^i \neq \sigma_P(\varphi^i)$. Then, because both $T$ and $T'$ have depth $J$ under Assumption (A1), the number of vertices for $G_{\varphi^i}$ and $G'_{\sigma_P(\varphi^i)}$ must be the same. Consequently, the order of $\varphi^i$ is different from $\sigma_P(\varphi^i)$.

Suppose that there exists some vertices $V_\ell$ such that for all $1 < \ell < J - 1$, $\varphi^i_\ell = \sigma_P(\varphi^i)_\ell$. Because each node has at least two children node due to Assumption (A4), then there exists a path, $\varphi''$, such that for all $1 \leq j < \ell + 1$, $\varphi^i_j = \varphi''_j$ and for all $\ell + 1 < j' \leq J$, $\varphi^i_j \neq \varphi''_{j'}$. Then, by definition of the path-preserving permutation, $\varphi''_{\ell+1} \neq \sigma_P(\varphi'')_\ell$ due to $\varphi''_{\ell+1} \neq \sigma_P(\varphi^i)_\ell$. Because all vertices are distinct, $\varphi''_{\ell+1} \not\in \varphi''_\ell, \sigma_P(\varphi'')_\ell$. Hence, $d_H(G_{\varphi''}, G'_{\sigma_P(\varphi'')}) \neq 0$.

Note that if such vertices do not exist, this argument can be repeated with the root node. As a result, we have found our contradiction.

$d_P(T, T') = d_P(T', T)$: By the properties of the Hausdorff metric, $d_H(G_{\varphi^i}, G'_{\sigma_P(\varphi^i)}) = d_H(G'_{\varphi^i}, G_{\sigma_P(\varphi^i)})$ so long as the vertices along the paths being compared are the same. Because $\sigma_P$ is
a bijection, define $\sigma_\mathcal{P}^{-1}$ to be the inverse. It immediately follows that $\arg\min_{\sigma_\mathcal{P}} \sum_z d_H(G'_z, G_{\sigma_\mathcal{P}(z)}) = \sigma_\mathcal{P}^{-1}$. As a result, the vertices along the paths are the same.

$$d_\mathcal{P}(T, T') \leq d_\mathcal{P}(T, T'') + d_\mathcal{P}(T', T'') :$$ Suppose not. Suppose there exists $T, T', T''$ such that

$$d_\mathcal{P}(T, T') > d_\mathcal{P}(T, T'') + d_\mathcal{P}(T'', T')$$

and $\sigma_\mathcal{P}$ is the bijection for $d_\mathcal{P}(T, T')$, $\sigma_\mathcal{P}'$ for $d_\mathcal{P}(T'', T')$, and $\sigma_\mathcal{P}''$ for $d_\mathcal{P}(T, T'')$. Then, because $\sigma_\mathcal{P}'$ is a bijection,

$$\sum_i d_H(G_{\varphi^i}, G_{\sigma_\mathcal{P}'(\varphi^i)}) \leq \sum_i d_H(G_{\varphi^i}, G_{\sigma_\mathcal{P}''(\varphi^i)}) + d_H(G_{\sigma_\mathcal{P}'(\varphi^i)}, G_{\sigma_\mathcal{P}''(\varphi^i)})$$

$$= d_\mathcal{P}(T, T'') + \sum_i d_H(G_{\sigma_\mathcal{P}'(\varphi^i)}, G_{\sigma_\mathcal{P}''(\varphi^i)})$$

$$= d_\mathcal{P}(T, T'') + \sum_i d_H(G_{\varphi}, G_{\sigma_\mathcal{P}'(\varphi^i)})$$

$$= d_\mathcal{P}(T, T'') + d_\mathcal{P}(T'', T')$$

$$< d_\mathcal{P}(T, T').$$

This is a contradiction because we have found a bijection that minimizes $\sum_z d_H(G_{\varphi^i}, G_{\sigma_\mathcal{P}'(\varphi^i)})$ better than $\sigma_\mathcal{P}'$.

**Lemma C.1.2.** The metric defined in 4.3.4 is a valid metric.

**Proof.** Because of assumptions (A2) and (A3), the tree permutation is a bijective mapping.

$$d_\mathcal{T}(T, T') = 0 \iff T = T' :$$ Note that $d_j(T, T') = 0$ if and only if all nodes are the same up to a permutation as the tree permutation is a bijective mapping.

$$d_\mathcal{T}(T, T') = d_\mathcal{T}(T, T') :$$ This follows because the permutation at each level is a bijection and we can invert the mapping.

$$d_\mathcal{T}(T, T') \leq d_\mathcal{T}(T, T'') + d_\mathcal{T}(T'', T') :$$ This follows by a similar reasoning as used in the proof of the previous metric. If there are such trees that violate the triangle inequality, we can show that there is a better minimum mapping than the one used for $d_\mathcal{T}(T, T')$ because the $L_2$ distance respects the triangle inequality and we can split it like the Hausdorff distance.

**Lemma C.1.3.** Suppose that $T$ and $T'$ are two trees that assumptions (A1) to (A4). Further, the trees have distinct topics and the number of children for each node except for those nodes at the leaves is at least two. Then,

$$d_\mathcal{P}(T, T') \leq d_\mathcal{T}(T, T').$$

**Proof.** The assumptions ensure that $d_\mathcal{P}(T, T')$ and $d_\mathcal{T}(T, T')$ are valid metrics. If $d_\mathcal{M}(G_z, G_{\sigma_\mathcal{P}(z)})$
denotes the minimum-matching distance defined in [53] between two polytopes, it follows that

\[ \min_{\sigma_p} \sum_i d_H(G_{\phi^i}, G_{\sigma_p(\phi^i)}) \leq \min_{\sigma_p} \sum_i d_M(G_{\phi^i}, G_{\sigma_p(\phi^i)}) . \]

Suppose \( \sigma_p = \arg \min_{\sigma_p} \sum_i d_H(G_{\phi^i}, G_{\sigma_p(\phi^i)}) \). Then, let \( K \) be the set such that for \( i = 1, 2, \ldots, |\Phi(T)| \),

\[ K_i = \{ k \in \phi^i, k' \in \sigma_p(\phi^i) | \arg \max_{k, k'} \min_{\theta_k, \theta'_{k'}} \| \theta_k - \theta'_{k'} \| \vee \max_{\theta_k, \theta'_{k'}} \| \theta_k - \theta'_{k'} \| \} . \]

Then, if \( (\sigma_T)^{-1} \) denotes the inverse subtree preserving permutation,

\[ \min_{\sigma_p} \sum_i d_M(G_{\phi^i}, G_{\sigma_p(\phi^i)}) \]

\[ \leq \sum_{k, k'} \max \left( \sum_i \| k \in \phi^i \|, \sum_i \| k' \in \sigma_p(\phi^i) \| \right) \left\| \theta_i - \theta'_{i} \right\| \]

\[ \leq \sum_{k, k' \in K} \max \left( \sum_i \| k \in \phi^i \|, \sum_i \| k' \in \sigma_p(\phi^i) \| \right) \]

\[ \min_{\sigma_T(V_k), \sigma_T(\ldots(V_k))} \left\| \theta_i - \theta'_{i} \right\| \min_{\sigma_T(V_k), \sigma_T(\ldots(V_k))} \left\| \theta'_{i} - \theta_{i}^{-1}(\sigma_T^{(\ldots(V_k)))} \right\| \]

\[ \leq \sum_{k, k' \in K} \left( \sum_i \| k \in \phi^i \| \right) \min_{\sigma_T(V_k), \sigma_T(\ldots(V_k))} \left\| \theta_i - \theta'_{i} \right\| \min_{\sigma_T(V_k), \sigma_T(\ldots(V_k))} \left\| \theta'_{i} - \theta_{i}^{-1}(\sigma_T^{(\ldots(V_k)))} \right\| \]

\[ \leq d_T(T, T') \]

\[ \square \]

C.1.2 Identifiability

Lemma C.1.4. Suppose that we have a tree, \( T \), such that all paths have length \( J > 1 \). Assume that \( (A8) \) holds. Then, the union of simplices defined by the tree have at least two exposed vertices.

Proof. Note that this is trivially true for \( J = 2 \). Because the root vertex must be shared and the collinear constraint, Assumption (A7), requires that vertices cannot lie on the same line, then if a vertex isn’t exposed, it must be equivalent to some other vertex. This violates the fact that the vertices are unique.
We now prove the case for $J > 2$ using induction. For the base case, assume that the tree only has two paths. Let $G_1$ and $G_2$ denote the respective simplices. Then, because of the tree, the collinear constraint, and the simplex constraint, the root vertex must be exposed. As a result, without loss of generality, if $G_1$ contains $J - 1$ vertices from $G_2$, then $G_2 \subset G_1$. Hence, all vertices of $G_1$ must be exposed and so there must be $J > 2$ vertices exposed.

Without loss of generality, if $G_1$ contains $J - 2$ vertices from $G_2$, then it can be shown that $G_2$ can only contain one exposed vertex from $G_1$. Let $\tilde{G}_2$ be the simplex defined by the contained $J - 2$ vertices and root. Then, $\tilde{G}_2 \subset G_1$. Note that because of the simplex constraint, removing any point reduces the dimension of the simplex. Hence, the dimensional difference between $\tilde{G}_2$ and $G_2$ must be one. Further, no two vertices of $G_1$ can be in the same affine dimension. As a result, because $\tilde{G}_2$ can only be extended in one affine direction, it is impossible for $G_2$ to capture more than one vertex of $G_1$.

Assume now that this is true for $n$ paths and we wish to prove it true for $n + 1$ paths. Because of the tree structure, all vertices on a path below the root must be connected to the root. Due to the collinear constraint, it is impossible for another vertex to be in the same direction from the root to the vertex or the vertex to the root. As a result, because of the simplicial constraint and how removing a point reduces the dimension, any vertex other than the root cannot be enclosed by the union of simplices and thus must be exposed. Hence, any shared exposed vertex other than the root can only be hidden by being contained in another simplex.

Then, if the new simplex contains all exposed points of the previous $n$ simplices, then its vertices must be exposed. Because it has at least $J$ vertex and $J > 2$, we have demonstrated the inductive hypothesis to be true in this case. On the other hand, if the vertices of the new simplex is contained by the union of the previous simplices, the inductive hypothesis still holds. Finally, even if the new simplex contains a subset of exposed points, but its vertices are not exposed, the argument used for the base case shows that there must be at least two exposed vertices for this to happen. After all, while the root vertex may be hidden, there must be at least two exposed vertices from the simplex defined by any other path and itself. Even if each simplex contributes one exposed vertices and another simplex hides the exposed vertex from the new simplex, we can repeat this argument to show that there must be at least two exposed vertices. Because there is a finite number of simplices, these exposed vertices can only be covered by exposing at least one more vertex.

\[\blacksquare\]

### C.1.3 Consistency

**Theorem C.1.1.** Suppose that we have two trees, $T$, and $T'$, such that assumptions (A1) and (A2) hold. Suppose also that $\min_i \pi_i > c_1$ for some constant $c_1 > 0$. Assume that for all $i \in 1, 2, \ldots, \mathcal{I}_p$,
as defined in [53], both \( p_{\eta|G_i} \) and \( p_{\eta|G'_i} \) and any restrictions are \( \alpha - \) regular densities on \( G, G' \). Further, \( G, G' \) have the thick body property discussed in [53]. Then, if \( V(p_{\eta|\mathcal{T}}|\mathcal{T}, p_{\eta|\mathcal{T}'}|\mathcal{T}') \to 0 \), \( d_{\text{ULH}}(\mathcal{T}, \mathcal{T}') \to 0 \).

**Proof.** To establish consistency, we first aim to establish an "inverse" bound. This can be done by following the argument of Theorem 5 from [53].

Let \( S_{[n]} = \{x_n\}_{n=1}^N \) and associate it with a vector \( \eta_i \) for each path from \( \mathcal{T} \) and \( \eta'_i \) for each path from \( \mathcal{T}' \) for \( i = 1, 2, \ldots, \mathcal{I}_p \). Then, for any constant \( \epsilon \), we have by triangle inequality and Cauchy-Schwartz for any \( d \in 1, 2, \ldots, D \),

\[
\left| \tilde{\eta}_d - \sum (\pi_i \eta_i)_d \right| \leq \sum |\pi_i (\tilde{\eta}_d - (\eta_i)_d)| \\
\leq \sum (\tilde{\eta}_d - (\eta_i)_d)^2 \\
\leq \sum |\tilde{\eta}_d - (\eta_i)_d|.
\]

As a result, by Hoeffding’s inequality, we have that for some \( i' \in 1, 2, \ldots, \mathcal{I}_p \),

\[
2D \exp\left(-2n\epsilon^2\right) \geq P_{S_{[n]}|\eta_1, \eta_2, \ldots, \eta_{\mathcal{I}_p}} \left( \max_d \left| \tilde{\eta}_d - \sum (\pi_i \eta_i)_d \right| \geq \epsilon \right) \\
\geq P_{S_{[n]}|\eta_1, \eta_2, \ldots, \eta_{\mathcal{I}_p}} \left( \max_d \sum |\tilde{\eta}_d - (\eta_i)_d| \geq \epsilon \right) \\
\geq P_{S_{[n]}|\eta_1, \eta_2, \ldots, \eta_{\mathcal{I}_p}} \left( \max_d \{|\tilde{\eta}_d - (\eta_i)_d| \geq \epsilon \right)_{i=1}^{\mathcal{I}_p} \\
= \sum_i \pi_i P_{S_{[n]}|\eta_i} \left( \max_d |\tilde{\eta}_d - (\eta_i)_d| \geq \epsilon \right) \\
\geq c_0 P_{S_{[n]}|\eta_i} \left( \max_d |\tilde{\eta}_d - (\eta_{i'})_d| \geq \epsilon \right).
\]

This implies that

\[
P_{\eta_{i'} \times S_{[n]}|\mathcal{T}}(\|\tilde{\eta} - \eta_{i'}\| \geq \epsilon) \\
\leq P_{\eta_{i'} \times S_{[n]}|\mathcal{T}}(\max_d |\tilde{\eta}_d - (\eta_i)_d| \geq \epsilon(D)^{-\frac{1}{2}}) \\
\leq \frac{2D}{c_0} \exp\left\{ \frac{-2n\epsilon^2}{D} \right\}.
\]

Then, define the event \( B = \{\|\tilde{\eta} - \eta_i\| \geq \epsilon\}_{i=1}^{\mathcal{I}_p} \) and \( B_i = \|\tilde{\eta} - \eta_i\| \geq \epsilon \) for \( i = 1, 2, \ldots, \mathcal{I}_p \). Due to assumption (A2), we can define a bijective permutation, \( \sigma \), between \( \Phi(\mathcal{T}) \) and \( \Phi(\mathcal{T}') \). It follows from our argument above that for any set \( A \),

\[
\left| P_{S_{[n]}|\mathcal{T}}(\tilde{\eta} \in A) - P_{S_{[n]}|\mathcal{T}', \tilde{\eta} \in A} \right|
\]
\[= \sum \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A) \]

\[= \sum \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A; B) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A; B) \]

\[+ \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A; B^c) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A; B^c) \]

\[= \sum \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A; B_i) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A; B_i) \]

\[+ \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A; B_i^c) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A; B_i^c) \]

\[\geq \sum \pi_i P_{\eta \times S[n]} G_i (\tilde{\eta} \in A; B_i) - \pi'_i P_{\eta \times S[n]} G'_i (\tilde{\eta} \in A; B_i) \]

\[- \frac{4D I_\rho (1 - c_0)}{c_0} \exp \left\{ \frac{-2n \epsilon^2}{D} \right\}.\]

As a result, the inverse bound can be established by finding a set such that only the simplices defined by paths from one tree are available. Then, without loss of generality, suppose that the union Hausdorff distance is defined between some simplex, \(G_i\), defined by path \(i\) from \(T\) and \(G'_i\), defined by path \(i'\) from \(T'\). In other words, \(d_H (G_i, G'_i) = d_H (T, T')\) for \(\eta \in G_i\) and \(\eta' \in G'_i\). Because of Assumption (A1), this is a meaningful comparison as the dimension of \(G_i\) and \(G'_i\) are of the same dimension for all \(i, i' \in 1, 2, \ldots, I_\rho\).

Then, there are a few scenarios to consider. First, suppose that \(G_i \cap G''_{i''} = \emptyset\) and \(G'_i \cap G''_{i''} = \emptyset\) for \(i'' \in 1, 2, \ldots, I_\rho\). Then, letting \(\epsilon_1 = \frac{d_H (T, T')}{4}\), we can clearly apply the argument of steps 4 and 5 of the proof of Theorem 5 from [53] on \(G_i, G'_i\) to derive an inverse bound.

Next, without loss of generality, suppose that there is a subset of paths, \(\phi (T')\), such that \(G_i \cap G''_{i''} \neq \emptyset\) for all \(i'' \in \phi (T')\). It is possible that \(i' \in \phi (T')\), but \(G''_{i''}\) only intersects with \(G_i\). Define the convex completion, \(C_G\), between \(G_i\) and \(\bigcup_{i'' \in \phi (T')} G''_{i''}\) to be the following:

- \(C_G \subset G_i\)
- \((G_i \setminus \bigcup_{i'' \in \phi (T')} G''_{i''}) \cup C_G\) is the union of convex sets
- \(|C_G|\) is minimized.

In other words, \(C_G\) re-partitions \(G_i \setminus \bigcup_{i'' \in \phi (T')} G''_{i''}\) into a union of polytopes by adding faces. Denote this union of polytopes as \(\bigcup B\). Note that our assumption on \(\alpha\)-regularity and on thick body means that any \(B\) must also have those properties. As \(C_G\) is the union of faces and by definition of the union Hausdorff distance, the distance must be computed between some \(B\) and \(G'_i\). We can also
apply the argument of steps 4 and 5 of the proof of Theorem 5 from [53] on B and \( G' \) as both are convex. As a result, we are again able to derive this inverse bound.

The last case is similar to case discussed above. However, there now also is a subset of paths, \( \phi(\mathcal{T}) \), such that \( G' \cap G'' \neq \emptyset \) for \( i'' \in \phi(\mathcal{T}) \). It is possible that \( i \in \phi(\mathcal{T}) \) as well. Nonetheless, we can still use a similar argument as above. In addition to the previous convex completion and resulting union of polytopes, let \( C_G \) represent the convex completion between \( G' \) and \( \bigcup_{i'' \in \phi(\mathcal{T})} G'' \) and \( \bigcup B' \) the subsequent union of polytopes. As before, \( B' \) must have the thick body property and \( \alpha \)-regularity. We can then apply the argument of steps 4 and 5 of the proof of Theorem 5 from [53] on some \( B \) and \( B' \) to derive an inverse bound.

Then, through this inverse bound, we have that as \( V(p_{S[n]|T}, p_{S[n]|T'}) \to 0 \), \( d_{\mathcal{UH}}(\mathcal{T}, \mathcal{T'}) \to 0 \).

### C.2 Algorithms

#### C.2.1 VB derivations

Rather than derive the entire lemma at once, we break it up and prove each piece. In addition, note that

\[
P(x_{m,n} \mid \zeta_{m,n}, z_m, \{\theta_k\}_{k=1}^K, \Phi(\mathcal{T})) = \prod_{k=1}^K \prod_{d=1}^D \Pi((\beta_k)_d) \Pi(\zeta_{m,n} = \ell(V_k)) \Pi(x_{m,n} = d). \tag{C.1}
\]

This will also help us derive our VB updates. Further, let \( \ell(V_k) \) denote the level of node \( V_k \).

First, let us derive updates for \( \beta_m \) and \( \theta_k \).

**Lemma C.2.1.** For \( \beta_m \),

\[
q(\beta_m) = \text{Dir}(\tilde{\alpha}_{\beta_m}),
\]

where for \( j \in 1, 2, \ldots, J \),

\[
(\tilde{\alpha}_{\beta_m})_j = (\alpha_{\beta})_j + \sum_{n=1}^N \mathbb{E}_{\zeta_{m,n}} (\Pi(\zeta_{m,n} = j)).
\]

**Proof.** From the decomposition in Equation 4.11, if we keep the terms that depend on \( \beta_m \), we have
that

$$\log(q(\beta_m)) \propto \mathbb{E}\left( \log \left( \prod_{n=1}^{N} P(\zeta_{m,n} | \beta_m) P(\beta_m | \alpha_\beta) \right) \right)$$

$$= \mathbb{E}\left( \sum_{n=1}^{N} \log(P(\zeta_{m,n} | \beta_m)) \right) + \log(P(\beta_m))$$

$$\propto \mathbb{E}\left( \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{I}(\zeta_{m,n} = j) \log((\beta_m)_j) \right) + \sum_{j=1}^{J} ((\alpha_\beta)_j - 1) \log((\beta_m)_j)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{J} \mathbb{E}_{\zeta_{m,n}} \left( \mathbb{I}(\zeta_{m,n} = j) \right) \log((\beta_m)_j) + \sum_{j=1}^{J} ((\alpha_\beta)_j - 1) \log((\beta_m)_j)$$

$$= \sum_{n=1}^{N} \sum_{j=1}^{J} \left( \mathbb{E}_{\zeta_{m,n}} \left( \mathbb{I}(\zeta_{m,n} = j) \right) \right) + (\alpha_\beta)_j - 1) \log((\beta_m)_j)$$

This is the kernel of a Dirichlet distribution with the parameter given in the lemma. \(\square\)

**Lemma C.2.2.** For \(\theta_k\),

$$q(\theta_k) = \text{Dir}(\bar{\alpha}_{\theta_k}),$$

where for \(d \in 1, 2, \ldots, D\) and \(l_k\) denotes the level for \(V_k\),

$$(\bar{\alpha}_{\theta_k})_d = (\alpha_\theta)_d + \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbb{I}(x_{m,n} = d) \mathbb{E}_{\zeta_{m,n}} \left( \mathbb{I}(\zeta_{m,n} = l_k) \right) \mathbb{E}_{z_m} \left( \mathbb{I}(k \in \varphi_{z_m}) \right).$$

**Proof.** From the decomposition in Equation 4.11, if we keep the terms that depend on \(\theta_k\) and repeat the argument from the previous lemma, we have that \(\log(q(\theta_k))\) is proportional to the following:

$$\mathbb{E}\left( \log \left( \prod_{m=1}^{M} \prod_{n=1}^{N} P(x_{m,n} | \beta_d, \zeta_{m,n}, \varphi_{z_m}) P(\theta_k | \alpha_\theta) \right) \right)$$

$$\propto \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{d=1}^{D} \mathbb{I}(x_{m,n} = d) \mathbb{E} \left( \mathbb{I}(\zeta_{m,n} = l_k) \mathbb{I}(k \in \varphi_{z_m}) \right) \log((\theta_k)_i) +$$

$$(\alpha_\theta - 1) \sum_{d=1}^{W} \log((\theta_k)_d)$$

$$= \sum_{d=1}^{D} \left( \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbb{I}(x_{m,n} = d) \mathbb{E}_{\zeta_{m,n}} \left( \mathbb{I}(\zeta_{m,n} = l_k) \right) \mathbb{E}_{z_m} \left( \mathbb{I}(k \in \varphi_{z_m}) \right) + (\alpha_\theta)_i - 1 \right) \log((\theta_k)_d).$$

This is the kernel of a Dirichlet distribution with the parameter given in the lemma. \(\square\)
Lemma C.2.3. For \( \pi \),

\[
q(\pi) = \text{Dir}(\tilde{\alpha}_\pi),
\]

where for \( i \in 1, 2, \ldots, |\Phi(T)| \),

\[
(\tilde{\alpha}_\pi)_i = (\alpha_\pi)_i + \sum_{m=1}^{M} \mathbb{E}(\mathbb{I}(z_m = i)).
\]

Proof. From the decomposition in Equation 4.11, if we keep the terms that depend on \( \theta_k \) and repeat the argument from the previous lemma, we have that

\[
\log(q(\pi)) \propto \mathbb{E} \left( \log \left( \prod_{m=1}^{M} P(z_m | \pi) P(\pi | \alpha_\pi) \right) \right)
\]

\[
= \mathbb{E} \left( \sum_{m=1}^{M} \log(P(z_m | \pi)) \right) + \log(P(\pi))
\]

\[
= \mathbb{E} \left( \sum_{m=1}^{M} \sum_{i=1}^{|\Phi(T)|} \mathbb{I}(z_m = i) \log((\pi_d)_i) \right) + \sum_{i=1}^{|\Phi(T)|} ((\alpha_\pi)_i - 1) \log((\pi_d)_i)
\]

\[
= \sum_{i=1}^{|\Phi(T)|} \left( \sum_{m=1}^{M} \mathbb{E}(\mathbb{I}(z_m = i)) + (\alpha_\pi)_i - 1 \right) \log((\pi_d)_i).
\]

This is the kernel of a Dirichlet distribution with the parameter given in the lemma. \( \square \)

Based on Lemma C.2.1 and C.2.2, we can derive updates for \( q(\zeta_{m,n}) \) and \( q(c_d) \).

Lemma C.2.4. For \( (\zeta_{m,n}) \) and \( j \in 1, 2, \ldots, J \),

\[
q(\zeta_{m,n} = j) \propto \exp \left( \psi((\tilde{\alpha}_\beta)_j) - \psi \left( \sum_{j'}(\tilde{\alpha}_\beta)_j' \right) + \right.

\[
\sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I}(x_{m,n} = d) \mathbb{I}(\zeta_{m,n} = \ell(V_k)) \mathbb{E}_{z_m}(\mathbb{I}(k \in \varphi^z_m))
\]

\[
\left. \left( \psi((\tilde{\alpha}_{\theta_k})_d) - \psi \left( \sum_{d'}(\tilde{\alpha}_{\theta_k})_d' \right) \right) \right),
\]

Proof. From the decomposition in Equation 4.11, if we keep the terms that depend on \( \zeta_{m,n} \) and use
Lemma C.2.5. For Lemma C.2.1 and C.2.2, we have that

\[
\log(q(\zeta_{m,n} = j)) \propto \mathbb{E} \left( \log(P(x_{m,n} | \beta_d, \zeta_{m,n}, \varphi^{z_m})P(\zeta_{m,n} = j | \beta_m)) \right)
\]

\[
= \mathbb{E} \left( \log(P(x_{m,n} | \beta_d, \zeta_{m,n} = j, \varphi^{z_m})) + \log(P(\zeta_{m,n} = j | \beta_m)) \right)
\]

\[
= \mathbb{E} \left( \log(\prod_{k=1}^{K} \prod_{d=1}^{D} ((\theta_k)_d)^{\mathbb{I}(x_{m,n} = d) \mathbb{I}(\ell(V_k) = j) \mathbb{I}(k \in \varphi^{z_m})}) \right) + \mathbb{E}_{\beta_m} \left( \log((\beta_m)_j) \right)
\]

\[
= \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{E} \left( \mathbb{I}(x_{m,n} = d) \mathbb{I}(\ell(V_k) = j) \mathbb{I}(k \in \varphi^{z_m}) \log((\theta_k)_d) \right) + \psi((\alpha_{\beta_m})_j) - \psi \left( \sum_{j'}(\alpha_{\beta_m})_{j'} \right)
\]

\[
= \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{E}_{\varphi^{z_m}} \left( \mathbb{I}(k \in \varphi^{z_m}) \mathbb{E}_{\theta_k} \left( \log((\theta_k)_i) \right) \right) + \psi((\alpha_{\beta_m})_j) - \psi \left( \sum_{j'}(\alpha_{\beta_m})_{j'} \right)
\]

\[
= \psi((\alpha_{\beta_m})_j) - \psi \left( \sum_{j'}(\alpha_{\beta_m})_{j'} \right)
\]

\[
\sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I}(x_{m,n} = d) \mathbb{I}(\zeta_{m,n} = \ell(V_k)) \mathbb{E}_{z_m} \left( \mathbb{I}(k \in \varphi^{z_m}) \right)
\]

\[
\left( \psi((\alpha_{\theta_k})_d) - \psi \left( \sum_{d'}(\alpha_{\theta_k})_{d'} \right) \right)
\].

\[\square\]

Lemma C.2.5. For \( z_m \) and \( i = 1, 2, \ldots, |\Phi(T)| \),

\[
q(z_m = i) \propto \exp \left( \psi((\alpha_{\pi})_i) - \psi \left( \sum_{i'}(\alpha_{\pi})_{i'} \right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I}(x_{m,n} = d) \mathbb{I}(k \in \varphi^i) \mathbb{E}_{\zeta_{m,n}} \left( \mathbb{I}(\zeta_{m,n} = \ell(V_k)) \right) \right)
\]

189
\[
\left( \psi((\bar{\alpha} d)) - \psi\left( \sum_{d'}(\bar{\alpha}_{d'}) \right) \right).
\]

**Proof.** From the decomposition in Equation 4.11, if we keep the terms that depend on \( \varphi^z \) and use Lemma C.2.1, C.2.2, and C.2.3, we have that

\[
\log(q(z_m = i)) \propto \mathbb{E} \left( \log \left( \prod_{n=1}^{N} P(x_{m,n} \mid \beta_m, \zeta_{m,n}, \varphi^i) P(i \mid \pi) \right) \right)
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{E} \left( \mathbb{I} (x_{m,n} = d) \mathbb{I} (l_k = \zeta_{m,n}) \mathbb{I} (k \in \varphi^i) \log((\theta_k)_d) \right) + \mathbb{E}_\pi \left( \log(P(\varphi^i \mid \pi)) \right)
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I} (x_{m,n} = d) \mathbb{I} (k \in \varphi^i) \mathbb{E}_\zeta (\mathbb{I} (l_k = \zeta_{m,n}) \log((\theta_k)_d)) + \psi((\bar{\alpha}_\pi)_i) - \psi\left( \sum_{i'}(\bar{\alpha}_\pi)_{i'} \right)
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I} (x_{m,n} = d) \mathbb{I} (k \in \varphi^i) \mathbb{E}_\zeta (\mathbb{I} (l_k = \zeta_{m,n})) \mathbb{E}_\theta (\log((\theta_k)_d)) + \psi((\bar{\alpha}_\pi)_i) - \psi\left( \sum_{i'}(\bar{\alpha}_\pi)_{i'} \right)
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I} (x_{m,n} = d) \mathbb{I} (k \in \varphi^i) \mathbb{E}_\zeta (\mathbb{I} (l_k = \zeta_{m,n})) \mathbb{E}_\theta (\log((\theta_k)_d)) + \psi((\bar{\alpha}_\pi)_i) - \psi\left( \sum_{i'}(\bar{\alpha}_\pi)_{i'} \right)
\]

\[
= \psi((\bar{\alpha}_\pi)_i) - \psi\left( \sum_{i'}(\bar{\alpha}_\pi)_{i'} \right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{d=1}^{D} \mathbb{I} (x_{m,n} = d) \mathbb{I} (k \in \varphi^i) \mathbb{E}_\zeta (\mathbb{I} (\zeta_{m,n} = \ell(V_k)))
\]

\[
\left( \psi((\bar{\alpha}_k)_d) - \psi\left( \sum_{d'}(\bar{\alpha}_k)_{d'} \right) \right).
\]

\[\square\]

**C.2.2 Geometric algorithm**

**C.2.2.1 Proof for single linkage merge groups**

**Corollary C.2.1.** Assume the assumptions of Lemma 4.4.5 hold. Then, the algorithm discussed above returns a tree set of merge groups, \( \Xi_T \). Further, if \( \tilde{\theta}_k = \sum_{(i,j) \in \Xi_T} \frac{1}{|\Xi_T|} \theta_{ij} \) for \( k = 1, 2, \ldots, K \), then \( \|\tilde{\theta}_k - \theta_k\| < \frac{\epsilon}{2} \).
Proof. By construction of the algorithm, an appropriate tree set of merge groups is discovered. Then, as discussed earlier, under our assumptions, \( \| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \| < \frac{\epsilon}{2} \) if \( \hat{\theta}_{i,j}, \hat{\theta}_{i',j'} \) are estimates of \( \theta_k \) for some \( k = 1, 2, \ldots, K \) and \( \| \hat{\theta}_{i,j} - \hat{\theta}_{i',j'} \| > \frac{\epsilon}{2} \) otherwise for all \( i \in 1, 2, \ldots, I_\rho \) and \( j = 1, 2, \ldots, J \). Our goal is to show that the only merge groups of the right size for node \( k \) that minimize \( \max_{(i,j),(i',j') \in M_k} \| \hat{\theta}_{(i,j)} - \hat{\theta}_{(i',j')} \| \) must be an appropriate merge group for that node.

We will do so by induction. In the base case, we examine the merge group for the root. Note that it cannot be the union of a smaller group with other indices because we can find some other node such that \( \| \hat{\theta}_{(i,j)} - \hat{\theta}_{(i',j')} \| > \frac{\epsilon}{2} \). As a result, there is only one valid merge group for the root.

Now, suppose all merge groups before node \( k \) have been appropriately assigned and we now seek to assign a merge group to node \( k + 1 \). Again, note that it cannot be the union of a smaller group with other indices due to the same logic from above. We can also rule out identifying a subset of a larger merge group. If the larger merge group corresponds to a node further up the tree or a node at the same level, that merge group has been identified earlier in our algorithm. As there is only one estimate of each simplices’ extreme points and the algorithm removes the indices of these estimates from consideration once assigned to a merge group, the merge group cannot be the subset of a larger merge group in this scenario. On the other hand, if the larger merge group corresponds to a node further below the tree, that node must be the child of some other node with a larger merge group. Because that node’s merge group has been identified, those groups are not under consideration. This situation can also be ruled out.

Then, if there are multiple merge groups of the same size available to be assigned, there must be nodes at the same level with equal number of paths that go through it. This can be seen by ruling out the alternatives. Merge groups for the parent nodes have already been assigned so it cannot be a merge group of a parent node. Further, by the same logic given in the previous paragraph, it cannot be the merge group corresponding to the child node of some other node.

As a result, the only ambiguous case is if there are multiple merge groups of the same size available to be assigned and multiple nodes at the same level with equal number of paths that go through it. However, because the algorithm tracks which groups are assigned to the parent node and the groups attached to the children nodes are a subset of the parent nodes’ groups, the subtrees are preserved. As a result, if \( T' \) is ultimately the tree estimated and \( T \) is the original tree, \( d_T(T, T') = 0 \). Hence, we can permute the labels of the topics according to \( T' \) without affecting the distance between the final estimates and the topic they’re trying to estimate. This establishes the inductive step. \( \square \)
C.2.2.2 Alternative merge group procedure

An alternative approach based on the proof of Lemma 4.4.5 is to convert the results from a k-means algorithm to the merge group. To discuss this approach, we first need to describe how to convert a cluster assignment to a merge group of size $O$ given a set of paths $\mathcal{I}$ and their collection of available extreme points indices, $\mathcal{J}_i$ for $i \in \mathcal{I}$.

- If $\xi_{i,j}$ denotes the label for $\widehat{\theta}_{i,j}$, let $\widehat{M}^{l} = \{(i, j) : \xi_{i,j} = l\}$.
- If $(\widehat{M}^{l})_{l=1}$ is identical for $o \in \tilde{O} \subset \{1, 2, \ldots, O\}$, keep $o_i$ such that $\max_{n, e \in o \cup \{1, 2, \ldots, S\}} \|\widehat{\theta}_{M^o_n} - \widehat{\theta}_{M^e_n}\|$ is minimized. If there are multiple such $o_i$, keep $o_i$ such that $\max \min_{i \in \mathcal{I}, j \in \mathcal{J}_i, (i,j) \notin \widehat{M}^{l}} \|\widehat{\theta}_{M^o_n} - \widehat{\theta}_{M^e_n}\|$ is minimized. If there are still multiple such $o_i$, then pick one at random.
- If $|\widehat{M}^{l}| < O$, use single linkage to add pairs to $\widehat{M}^{l}$ so that it is a merge group of size $O$ for a set of paths $\mathcal{I}$ and their collection of available extreme points indices, $\mathcal{J}_i$ for $i \in \mathcal{I}$.
- If $|\widehat{M}^{l}| > O$, remove $|\widehat{M}^{l}| - O$ pairs one at a time such that $\max_{o, o' \in \widehat{M}^{l}} \|\widehat{\theta}_{M^o_n} - \widehat{\theta}_{M^o_n}\|$ is minimized as much as possible.

Then, we can use this procedure for the following alternative algorithm.

1. Perform k-means on $\{\widehat{\theta}_{1,j}\}_{j=1}^I, \{\widehat{\theta}_{2,j}\}_{j=1}^I, \ldots, \{\widehat{\theta}_{1,\mathcal{J}_o}\}_{j=1}^I$ with $K$ as the number of clusters. Let $\zeta_{i,j}$ denote the cluster assignment for $\widehat{\theta}_{i,j}$. Let $\tilde{J}_i = \{1, 2, \ldots, J\}$ for $i = 1, 2, \ldots, \mathcal{J}_o$.

2. Suppose $\widehat{M}^{l}$ is the largest group. Convert $\widehat{M}^{l}$ to a merge group, $\mathcal{M}^{l}$, of size $\mathcal{I}_o$ for $\Phi(\mathcal{T})$ and their $\tilde{J}_i$. If any pairs were added to the merge group, set their $\zeta_{i,j} = k$. If any pair, $(i, j)$, are removed from $\widehat{M}^{l}$ in its conversion, set $z_{i,j} = l'$ where $l'$ is the closest assignment and $\widehat{M}^{l'}$ has not been converted to a merge group. If there are multiple such groups, convert them all to a merge group and choose the one with the smallest $\max_{(i,j),(i',j')} \|\widehat{\theta}_{(i,j)} - \widehat{\theta}_{(i',j')}\|$. If $(i, j) \in \mathcal{M}^{l}$, remove $j$ from $\tilde{J}_i$.

3. In order of the number of paths that go through a node at the second level, repeat the previous step for the vertices at that level using the largest remaining assignments. The size of the merge group is the number of paths that go through it. Further, if $(i, j) \in \mathcal{M}^k$ for some node $V_k$, then $(i, j) \notin \mathcal{M}^{k'}$ for some other node at the second level, $V_{k'}$. Finally, let $\Phi^k(\mathcal{T}) = \{i : i \in \mathcal{M}^k\}$ for some $V_k$ at the second level.

4. Repeat step 2 for the vertices at the next level. However, the set of paths is restricted to $\Phi^k(\mathcal{T})$ where $V_k$ is the parent of the node under consideration. If there are no such
paths, consider $\tilde{M}^j$ such that $\sum_{i \in \tilde{M}^j} \mathbb{I} (i \in \Phi^k(\mathcal{T}))$ is maximized. Again, if there are multiple groups, convert them all to a merge group and select the one with the smallest $\max_{(i,j), (i',j') \in M^{k'}} \left\| \hat{\theta}_{ij} - \hat{\theta}_{i'j'} \right\|$. 

5. Repeat this for the children’s childrens and so on until merge groups at level $J - 1$ have been identified. The unassigned vertices are associated with the nodes at the leaf.

6. Return $\overline{\theta}_k = \sum_{M^k} \frac{1}{|M^k|} \hat{\theta}_{i,j}$ for $k \in 1, 2, \ldots, K$.

Under the conditions specified in the lemma 4.4.5, this algorithm will return the optimal result because the estimates closest to each other are estimates of the same topic. However, because we are uncertain of whether we are in that regime, we need to modify the k-means results so that they are a tree set of merge groups.

C.2.2.3 Hybrid geometric algorithm

For this algorithm, we modify VLAD in the following manner [75]:

1. Given $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_M$, set $\bar{\eta}_0 = \frac{1}{M} \sum \hat{\eta}_m$ and $\hat{\eta}_m = \hat{\eta}_m - \bar{\eta}_0$ for $m = 1, 2, \ldots, M$.

2. Let $E \in \mathbb{R}^{M \times D}$ such that $E_{m, \cdot} = \hat{\eta}_m$. Compute the top $J - 1$ singular factors of $E$, i.e. $E = U\lambda W^T$.

3. Set $\varsigma_1, \varsigma_2, \ldots, \varsigma_J$ to be the K-means centers of $u_1, u_2, \ldots, u_M$. Here, $u_m$ is a row of $U$.

4. For $j = 1, 2, \ldots, J$, set $\hat{\theta}_j = \gamma_j W\Lambda \varsigma_j + \bar{\eta}_0$. Let $G$ be the simplex defined by these vertices and $\mathcal{T}$ the tree associated with $G$. Find $\gamma_1, \gamma_2, \ldots, \gamma_J \in \mathbb{R}^+$ so that

- $\sum d_\lambda(\hat{\eta}_m, \mathcal{T}, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_J)$ is minimized and
- $\text{vol}(G)$ is minimized.

5. Return $\{\hat{\theta}_j\}_{j=1}^J$. 

193
Table C.1: Time in seconds that it took to run the various approaches for data of different dimensions. For the Gibbs sampler, 4 chains with 2000 iterations were run in parallel.

<table>
<thead>
<tr>
<th>Method</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA Gibbs</td>
<td>138.093</td>
<td>144.956</td>
<td>142.221</td>
<td>143.337</td>
</tr>
<tr>
<td>Tree Gibbs cf (4.4.2)</td>
<td>37274.930</td>
<td>40711.527</td>
<td>38001.576</td>
<td>40921.788</td>
</tr>
<tr>
<td>Tree VB cf (4.4.3)</td>
<td>2809.124</td>
<td>2603.870</td>
<td>2676.995</td>
<td>2934.226</td>
</tr>
<tr>
<td>GeoTree (VLAD) cf [75]</td>
<td>23.611</td>
<td>20.136</td>
<td>18.867</td>
<td>18.183</td>
</tr>
<tr>
<td>GeoTree (GDM) cf [76]</td>
<td>89.796</td>
<td>79.505</td>
<td>55.447</td>
<td>79.285</td>
</tr>
<tr>
<td>GeoTree (Hybrid) cf (C.2.2.3)</td>
<td>80.158</td>
<td>64.975</td>
<td>49.464</td>
<td>80.542</td>
</tr>
</tbody>
</table>

Table C.2: Time in seconds that it took to run the various approaches for different vocabulary sizes. For the Gibbs sampler, 4 chains with 2000 iterations were run in parallel.

<table>
<thead>
<tr>
<th>Method</th>
<th>32</th>
<th>34</th>
<th>52</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA Gibbs</td>
<td>566.113</td>
<td>1297.704</td>
<td>1868.530</td>
</tr>
<tr>
<td>Tree Gibbs</td>
<td>128459.296</td>
<td>117150.381</td>
<td>237127.254</td>
</tr>
<tr>
<td>Tree VB</td>
<td>5645.118</td>
<td>6107.874</td>
<td>9900.026</td>
</tr>
<tr>
<td>GeoTree (VLAD)</td>
<td>169.144</td>
<td>213.820</td>
<td>213.509</td>
</tr>
<tr>
<td>GeoTree (GDM)</td>
<td>246.052</td>
<td>278.420</td>
<td>452.160</td>
</tr>
<tr>
<td>GeoTree (Hybrid)</td>
<td>233.399</td>
<td>242.236</td>
<td>687.234</td>
</tr>
</tbody>
</table>
Table C.3: Time in seconds that it took to run the various approaches for different vocabulary sizes. For the geometric approaches, the times were averaged across five runs.

### C.3 Simulation results

#### C.3.1 Algorithm timing

### C.4 Real data results

#### C.4.1 Algorithm timing

#### C.4.2 Exploration of GeoTree result

Figure C.1 shows the topic tree, i.e. how each topic is organized in the fitted tree and the five most likely words associated with each topic for a tree of depth 4 and each node except the leaf nodes having three children. Of the five runs, this estimated tree had the smallest projected simplicial distance. There are a few points that we would like to discuss. First, the top five words for the topic at the root are related to people, time, or place. This makes sense because after removing the stop words, news articles are usually about what people do at a given time point and place. On the other hand, the topics at the bottom of the tree are more specific. Topic 14 appears to be about art museum whereas Topic 23 clearly refers to the United Nations. This is to be expected because more common words are pushed to the top of the tree.

Next, it is worth understanding the topic simplices defined by the paths because it might suggest a certain type for a subset of articles. For instance, the path connecting topics 1, 2, 5, and 14 might be about exhibitions at art museums because Topic 5 are about tickets and includes locations and as discussed previously, Topic 14 is on art museum. As another example, the path connecting 1, 4, 12, and 36 suggests that there is a subset of articles related to the Florida recount during the 2000 United States presidential election. Topic 36 appears related to the courts’ review of state election law whereas Topic 4 and 12 includes the word "Bush", "Gore", and "Florida".

Finally, comprehending how the topics are arranged is worth discussing because of the relationships it reveals. For instance, the subtree of Topics 1, 4, 12, and the children nodes of 12 appear
Figure C.1: Best fitted topic tree of the five runs showing the five most likely words associated with each topic from various methods according to the projected simplicial distance. The number of documents associated with each path is also displayed at the bottom of the tree. The left one is derived from geometric method whereas the right one is from VB.
to be about the presidential candidates. Specifically, it is about their popularity and Bush’s tax plan. On the other hand, these relationships demonstrate how words might be used in different contexts. For instance, under Topic 2 is Topic 6. Topic 6 features the words "house", "Israel", and "peace". The word "house" is tied to theater from Topic 2 because shows are performed at theater houses. Indeed, Topic 18 is a topic about musicals. However, the word "house" might also refer to the White House. In particular, the Camp David Summit was held in 2000 in hopes of ending the Israeli-Palestine conflict. This obviously would have involved the president and the White House and explains some of the words associated with Topic 6 and 17. As seen in Topic 19, the White House is also located in Washington DC, which has a lot of office buildings. Because similar such arguments can be made for the layout of topics elsewhere in the tree, it is encouraging that the geometric methods are able to discover these semantic relationships.

It is also interesting to compare the topic tree learned by the geometric algorithms against the topic tree learned by VB. While the top three words are the same for the root topics, the other topics are different. Even though the topics are logically consistent, the VB topics are much less specific than the geometric algorithm topics regardless of the level. For instance, the top five words associated with Topic 39, which is at the bottom of the VB tree, are "music", "film", "theater", "play", and "show". On the other hand, the geometric tree has a musical topic, Topic 18, and a film festival topic, Topic 25, at the bottom of its tree. As a result, the subjects defined by the paths from the VB tree are more general. For example, there is no explicit mention of the Florida recount during the elections. Instead, all election news is primarily folded into the path of topics 1, 3, 8 and 24 because the top five words associated with these topics include president, Bush, Gore, and campaign and there are 1,449 articles associated with that path. Meanwhile, the topic referencing the Supreme Court or Topic 29 is on a separate path, i.e the path of topics 1, 3, 10, and 29. While the first two topics are the same, Topic 10 appears to be about the criminal justice system. As a result, it is hard to tease out the Florida recount. This example also demonstrates that because the subjects from the VB tree are more general, they are perhaps easier to identify and differentiate. Finally, the path based metric and tree based metric between the mean posterior VB tree based mixture and the geometric fitted tree based mixture are 3.833 and 9.970 respectively. On the other hand, the average projected simplicial distance for the mean posterior VB tree based mixture is 0.112, which is minutely better than that of the projected simplecial distance. This perhaps suggest that these two methods provide different, but similarly reasonable explanations for the corpus. Alternatively, if we examine the topics from the geometric methods, we see that perhaps topic 34 should belong under topic 3 because both include game and team. However, because the vertices of topic 34’s probability simplex weren’t as "close" as those of other probability simplices, it got put under topic 4 instead due to the geometric methods’ hard cutoff. In other words, it is possible that the number of children can vary by node or the length of the path can vary. VB might be better able to handle
this because it uses a soft cut-off when updating the estimates for each topics’ alpha prior.
BIBLIOGRAPHY


[34] Robert Hunt. Worked example: Geodesics on the surface of a sphere.


[57] Vera Pawlowsky-Glahn, Juan Jose Egozcue, and Raimon Tolosana-Delgado. Lecture notes on compositional data analysis. 2007.


