BAYESIAN NONPARAMETRIC METHODS FOR PROTEIN STRUCTURE PREDICTION

A Dissertation

by

KRISTIN PATRICIA LENNOX

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Statistics

BAYESIAN NONPARAMETRIC METHODS FOR PROTEIN STRUCTURE PREDICTION

A Dissertation

by

KRISTIN PATRICIA LENNOX

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	David B. Dahl
Committee Members,	Nancy Amato
	Raymond J. Carroll
	Thomas E. Wehrly
Head of Department,	Simon J. Sheather

August 2010

Major Subject: Statistics

ABSTRACT

Bayesian Nonparametric Methods for Protein Structure Prediction.

(August 2010)

Kristin Patricia Lennox, B.S., Texas A&M University;

M.S., Texas A&M University

Chair of Advisory Committee: Dr. David B. Dahl

The protein structure prediction problem consists of determining a protein's three-dimensional structure from the underlying sequence of amino acids. A standard approach for predicting such structures is to conduct a stochastic search of conformation space in an attempt to find a conformation that optimizes a scoring function. For one subclass of prediction protocols, called template-based modeling, a new protein is suspected to be structurally similar to other proteins with known structure. The solved related proteins may be used to guide the search of protein structure space.

There are many potential applications for statistics in this area, ranging from the development of structure scores to improving search algorithms. This dissertation focuses on strategies for improving structure predictions by incorporating information about closely related "template" protein structures into searches of protein conformation space. This is accomplished by generating density estimates on conformation space via various simplifications of structure models. By concentrating a search for good structure conformations in areas that are inhabited by similar proteins, we improve the efficiency of our search and increase the chances of finding a low-energy structure.

In the course of addressing this structural biology problem, we present a number of advances to the field of Bayesian nonparametric density estimation. We first develop a method for density estimation with bivariate angular data that has applications to characterizing protein backbone conformation space. We then extend this model to account for multiple angle pairs, thereby addressing the problem of modeling protein regions instead of single sequence positions. In the course of this analysis we incorporate an informative prior into our nonparametric density estimate and find that this significantly improves performance for protein loop prediction. The final piece of our structure prediction strategy is to connect side-chain locations to our torsion angle representation of the protein backbone. We accomplish this by using a Bayesian nonparametric model for dependence that can link together two or more multivariate marginals distributions. In addition to its application for our angular-linear data distribution, this dependence model can serve as an alternative to nonparametric copula methods.

For my parents

ACKNOWLEDGEMENTS

I wish to thank my advisor, David Dahl, for giving me the opportunity to undertake this project and for his invaluable guidance throughout. This work would not have been possible without the contributions of my collaborators, Marina Vannucci, Ryan Day, and Jerry Tsai, and they have my sincere gratitude. I would also like to thank my committee members, Nancy Amato, Ray Carroll, and Tom Wehrly, for their valuable time and advice. Last but far from least, I am deeply grateful to my family and friends for all of their support over the years.

While performing this research I was supported by NIH/NIGMS grant R01GM81631.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

1.1 Protein Structure Prediction

The explosion of bioinformatics data from sources such as the Human Genome Project has lead to something of an embarrassment of riches from the proteomics standpoint. While protein sequences based on genetic information are becoming increasingly available, the three-dimensional structures of these novel proteins remain elusive. These structures are of great importance to molecular biologists as they provide insights into the behavior of proteins in biological systems. The methods for experimentally determining protein structure, X-ray crystallography and NMR spectroscopy (Schlick, 2006, pp. 16–19), are both expensive and time consuming. Although advances have greatly increased the speed and affordability of these techniques (see e.g. Usón and Sheldrick, 1999; Clore and Schwieters, 2002) currently there is no experimental structure determination method which can keep pace with the genomics revolution.

This has lead to the rise of the field of protein structure prediction. Protein structure is generally divided up into four categories: primary through quaternary. Primary structure is the sequence of amino acids composing a protein, while secondary structure consists of well-defined three-dimensional motifs induced by hydrogen bonding. Tertiary structure describes how these regular elements fit together, giving the full three-dimensional structure of a polypeptide chain, while quaternary structure describes how multiple chains fit together. The typical goal of structure prediction is to arrive at tertiary structure using only

The format and style follow that of *Biometrics*.

primary structure as a starting point.

Methods for structure prediction rely on a guided search of conformation space. A typical example is the Rosetta software package (Das and Baker, 2008), which iterates between random perturbations of a protein's structure and the scoring of those changes using an approximate energy function. An open problem for statisticians working in this field is how to best integrate known information about protein structure into a probabilistic search of protein conformation space.

1.1.1 Template-Based Structure Prediction

Template-based modeling, sometimes called homology modeling, uses solved structures for similar proteins as a starting point when modeling a novel structure. Similarity is generally measured by sequence identity: the percentage of identical amino acids for each sequence position in two aligned proteins. As libraries of solved structures, such as the Protein Data Bank (PDB) (Kouranov et al., 2006), continue to grow, template-based modeling can be extended to new classes of proteins, and existing models can be refined by the addition of new data. The challenge for modern methods is to provide improvements over the closest match among the template proteins (Kopp et al., 2007).

Since homology modeling gives us a well-defined population of structures, it makes sense to discuss the distribution of such populations. A novel protein with high sequence identity to a known structure family can be treated as a draw from some distribution for these related structures. We can therefore base our conformation search on a distribution estimated using the solved template-family members. A natural method for developing such estimates is some form of nonparametric density estimation. This dissertation focuses on developing and applying techniques from Bayesian nonparametric density estimation to the problem of template-based modeling.

1.1.2 Bayesian Nonparametrics

The proposed statistical models for proteomics data stem from Bayesian nonparametric methods. As for all Bayesian inference, Bayesian nonparametrics methods require a probability distribution for all model parameters, making the name somewhat misleading. What distinguishes these methods is that models are indexed by infinite parameter spaces, rather than the finite parameter models of traditional parametric inference. Müller and Quintana (2004) provide an overview of the field.

While Bayesian nonparametric techniques can be applied to a wide variety of problems, in the context of structure prediction we are primarily interested in density estimation. Similar to frequentist kernel density estimation, these models can be viewed as infinite mixtures of (typically) unimodal distributions.

The workhorse of Bayesian nonparametric density estimation is the Dirichlet process first described by Ferguson (1973). The Dirichlet process is a distribution on almost surely discrete probability distributions. Using the stick-breaking representation of Sethuraman (1994), a random measure G drawn from a Dirichlet process $DP(\tau_0 G_0)$ takes the form:

$$G(\boldsymbol{B}) = \sum_{j=1}^{\infty} p_j \delta_{\gamma_j}(\boldsymbol{B})$$

where δ_τ is an indicator function equal to 1 if $\gamma \in \boldsymbol{B}$ and 0 otherwise, $\gamma_j \sim G_0$, $p'_j \sim$Beta$(1, \tau_0)$, $p_1 = p'_1$, and $p_j = p'_j \prod_{k=1}^{j-1}(1 - p'_k)$ for $j > 1$. The distribution G_0 is typically referred to as the centering distribution, and defines the general "shape" of the distributions drawn from the process. The parameter τ_0 is known as the mass parameter, and determines how concentrated the weights of the distribution G will tend to be.

The discreteness of draws from a Dirichlet process renders it unattractive for directly modeling continuous distributions. However Antoniak (1974) proposed the use of the Dirichlet process to define an infinite mixture model. Consider a set of observations

$x_1, ..., x_n$, and a family of distributions $f(x|\theta)$ indexed by parameter θ. A Dirichlet process mixture (DPM) model takes the form:

$$\begin{aligned} x_i \mid \theta_i &\sim f(x_i|\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim DP(\tau_0 G_0). \end{aligned}$$

The discreteness of G means that there is positive probability that $\theta_i = \theta_j$ for all i, j. Since equality of the parameters would imply x_i and x_j are drawn from identical distributions, they are considered to be clustered together. Hence, we are discussing a kind of mixture model with a draw from a Dirichlet process serving as a prior on mixture components. Note that some authors follow the convention of Antoniak in referring to such models as mixture of Dirichlet process models.

A Bayesian density estimate $P(x)$ based on such a model can be described by the predictive distribution:

$$P(x_{n+1}|x_1, ..., x_n) = \int f(x_{n+1}|\theta) dP(\theta|x_1, ..., x_n) \tag{1.1}$$

Density estimates of this form are discussed by Ferguson (1983) for the univariate case and by Tiwari et al. (1988) for multivariate densities. Lo (1984) provides an interpretation of such models in terms of a convolution between a kernel f and a Dirichlet process.

The density (1.1) generally can not be directly evaluated, and so some form of approximation must be used. Early work with DPM models, for density estimation and otherwise, was hampered by a lack of computational resources. Considerable effort was expended to explore approximations and algorithms for faster computation (see e.g. Berry and Christensen, 1979; Kuo, 1986; West, 1990). Advances in both computing power and Markov chain Monte Carlo (MCMC) techniques opened the door to sophisticated models for larger and more complex datasets. Of particular interest is a result from Escobar and West (1995).

They show that if one can generate B draws $\theta^{(1)}, ..., \theta^{(B)}$ from the posterior distribution $\theta | x_1, ..., x_n$ through some MCMC scheme, then the approximation given by

$$\hat{P}(x_{n+1}|x_1, ..., x_n) = \frac{1}{B} \sum_{i=1}^{B} f(x_{n+1}|\theta^{(i)})$$

is almost surely consistent for (1.1). Combined with work such as that of MacEachern and Müller (1998) and Neal (2000) on MCMC methods for sampling from the posterior distribution of DPM models with nonconjugate centering distributions, this gives a framework for density estimation beyond the realm of Gaussian mixtures.

The remainder of this dissertation will present a number of such extensions within the DPM framework. While these methods, ranging from the accommodation of angular data to the nonparametric modeling of complex association, were developed to address particular problems in protein structure prediction, they also have more general applications in Bayesian parametric and nonparametric statistics.

1.2 Modeling ϕ, ψ Angles

Chapter II develops a Bayesian model for bivariate angular data that is particularly useful in the context of protein structure prediction. A protein consists of a chain of covalently linked amino acids. Each amino acid has four heavy atoms (C, C_α, N, and O) in addition to a unique side-chain group. The side-chains are what distinguish the 20 unique amino acids from one another, while the four heavy atoms make up the so-called protein backbone (Schlick, 2006, p. 66).

A standard first step in protein structure prediction is generating candidate conformations for the protein backbone. This leads to difficulties with dimensionality; a full description of the backbone consists of the (x, y, z) coordinates for each heavy atom, which would mean a $12n$ dimensional space for a polypeptide consisting of n amino acids. However, the stereochemistry of the protein backbone leads to a simplification in the form of torsion angles. Ramachandran et al. (1963) noted that the backbone can be represented by a (ϕ, ψ) angle pair at each sequence position, reducing the dimensionality of the problem to $2n$. Unlike Cartesian coordinates, which can be modeled using standard linear data techniques, (ϕ, ψ) pairs require methods specific to angular data. (In particular, they exhibit wrapping: the property that $\phi = \phi + 2\pi k$ for any integer k.)

Chapter II focuses on the use of Bayesian nonparametric density estimation for bivariate angular data, and particularly density estimation for (ϕ, ψ) angles. It contains the necessary full conditional distributions and computational methods for Bayesian modeling using the bivariate von Mises sine model (Singh et al., 2002), which is a distribution that accounts for the wrapping of bivariate angular data. However, as the sine model is an elliptical distribution analogous to a bivariate normal, it cannot fully capture the behavior of torsion angles. Therefore a DPM model, which allows us to generate distributions incorporating multiple elliptical components, is more appropriate than a parametric model.

This technique is applied to assessing the use of “whole” versus “half” positions for template-based structure prediction. Whole position data consists of a (ϕ, ψ) angle pair for each sequence position, while a half position has a ψ angle from one position and a ϕ angle from the subsequent position. Because two sequence positions are involved, half positions can be classified into more specific categories by factors like amino acid and secondary structure type. In theory, this should allow for improved structure prediction. We demonstrate that this is indeed the case using data from the globin protein family.

1.3 Joint Distributions for the Protein Backbone

Chapter III presents models that can account for multiple (ϕ, ψ) pairs simultaneously. Recall that proteins consist of long chains of amino acids, and thus the backbone representation is a chain of (ϕ, ψ) angles. In order to adequately account for the structure of the data some kind of joint modeling is necessary. In the context of template-based modeling, such joint distributions are particularly useful for loop and turn regions in proteins.

The core of a protein is composed of regular secondary structure elements such as α-helices and β-strands. Such regions have highly conserved (ϕ, ψ) angle pairs, and are relatively easy to predict. However, these regular secondary structure regions are connected by flexible loops and turns. These areas can differ radically between members of the same protein family, and are also the regions most prone to amino acid insertions and deletions. Current knowledge-based loop prediction methods are either based on draws from coil libraries (e.g. Fitzkee et al., 2005) or on datasets that are not limited to proteins similar to the target (Boomsma et al., 2008).

Chapter III proposes a joint model for multiple (ϕ, ψ) pairs which is suitable for loop modeling and provides continuous density estimates. It also contains two prior formulations, the first of which is a direct extension of the single position model in Chapter II. The second is a Dirichlet process mixture of hidden Markov models (DPM-HMM). One characteristic of Bayesian density estimation not shared by kernel density methods is the ability to incorporate prior knowledge into the shape of the final density estimate, and the DPM-HMM takes full advantage of this property. The DPM-HMM infers the secondary structure type at each sequence position via a hidden Markov model as an intermediate step in density estimation. This significantly improves our model, as we can produce useful secondary structure based density estimates even at alignment positions with little or no observed data.

The necessary computational techniques and distributions for the use of this strategy are presented, along with a method for dealing with the problem of “sparse data” arising from amino acid insertions and deletions. The noninformative prior and DPM-HMM models are then compared to both the coil library of Fitzkee et al. (2005) and the *de novo* DBN-torus model of Boomsma et al. (2008) for the EF loop of the globin protein family.

1.4 Linking the Backbone to the Side-Chains

The challenge addressed in Chapter IV is how to link predictions for side-chain placement to those for the (ϕ, ψ) representation of the protein backbone. Modeling the protein backbone is crucial for protein structure prediction, but is not sufficient to construct a complete candidate structure. The placement of the amino acid side-chains, the residues branching off from the backbone which are unique to each amino acid, is also required. Although a typical side-chain consists of many atoms, a simplified representation is the position of the side-chain centroid.

When performing joint modeling for side-chain and backbone data, each observation consists of a (ϕ, ψ) pair for the backbone and a set of (x, y, z) coordinates for the side-chain centroid. This is a combination of angular and linear data types. Although joint distributions exist for angular and linear variables (Johnson and Wehrly, 1978), they are difficult to use in the mixture modeling context.

This situation is addressed as a special case of a more general problem: how does one develop a model of association between variables from different distribution families? This could refer to angular-linear combinations, categorical-numerical combinations, or any other situation where a standard multivariate distribution is unavailable or unsuitable.

The proposed model, referred to as a Dirichlet process dependence (DPD) model, can be used in any situation when component variables can be modeled separately with standard Bayesian procedures, but where a suitable joint model is nonobvious. Rather than explicitly defining correlation style parameters, a DPD model handles association exclusively through Dirichlet process induced clustering. This model can identify both distinct data populations and within population association. Note that the noninformative prior model for the protein backbone developed in Chapter III is an example of a DPD style model for association.

We apply this model to studying the relationship between side-chain and backbone conformations in protein cliques. A clique is a set of amino acids which are in close contact when a protein is folded, but not necessarily adjacent on the backbone. By investigating the relationship between clique residues and backbone conformations, we can determine more efficient methods for joint modeling and potentially develop side-chain driven structure prediction methods.

1.5 Conclusions

The final chapter of the dissertation addresses all of the previous work in the dual contexts of protein structure prediction and statistics. The previous chapters are examined both in terms of the advances they represent in structural biology and their contribution in the field of Bayesian statistical modeling. In addition, we discuss the roles of the various methods in an algorithm developed for use in the CASP (Moult, 2005) experiment for protein structure prediction. While the applications for these methods have been presented independently, they can be incorporated into a coherent strategy for protein structure prediction. Furthermore the methods described are by no means limited to the structural biology framework. Angular data is certainly not confined to biological applications, and DP models of association are also valuable to a wider audience. By examining these methods in a wider context, this section aims to give a more complete picture of the impact of this work.

CHAPTER II

DENSITY ESTIMATION FOR PROTEIN CONFORMATION ANGLES USING A BIVARIATE VON MISES DISTRIBUTION AND BAYESIAN NONPARAMETRICS*

2.1 Introduction

Computational structural genomics has emerged as a powerful tool for better understanding protein structure and function using the wealth of data from ongoing genome projects. One active area of research is the prediction of a protein's structure, particularly its backbone, from its underlying amino acid sequence (Dill et al., 2007).

Based on the fundamental work of Ramachandran (Ramachandran et al., 1963), the description of the protein backbone has been simplified by replacing the (x, y, z) coordinates of an amino acid residue's four heavy atoms (N, C_α, C, and O) with the backbone torsion angle pair (ϕ, ψ) (Figure 1). A standard visual representation is the Ramachandran plot, in which ϕ angles are plotted against ψ angles. Because of their importance to structure prediction and their simple representation, a great deal of recent work has sought to characterize the distributions of these angle pairs, with an eye towards predicting conformational angles for novel proteins (Ho et al., 2003; Xue et al., 2008).

Datasets from the Protein Data Bank (PDB) (Berman et al., 2003) can consist of over ten thousand angle pairs, which provide ample data for even relatively unsophisticated den-

* Material in this chapter is reprinted with permission from "Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics" by Lennox, K. P., Dahl, D. B., Vannucci, M. and Tsai J. 2009. *Journal of the American Statistical Association.* **104**, pp. 586–596, and "Correction to Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics" by Lennox, K. P., Dahl, D. B., Vannucci, M. and Tsai J. 2009. *Journal of the American Statistical Association.* **104**, p. 1728
Journal URL: http://pubs.amstat.org/loi/jasa

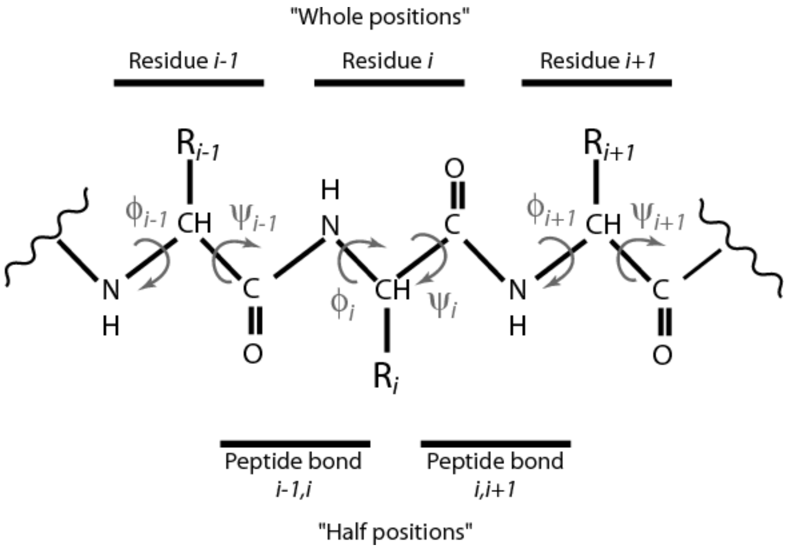


Figure 1: Diagram of protein backbone, including ϕ and ψ angles, whole positions, and half positions. At the ith residue, the ϕ angle describes the torsion around the bond N$_i$-C$_{\alpha i}$, measuring the angle between the C$_{i-1}$-N$_i$ and the C$_{\alpha i}$-C$_i$ bonds, while the ψ angle describes the torsion around the bond C$_{\alpha i}$-C$_i$, measuring the angle between the N$_i$-C$_{\alpha i}$ and the C$_i$-N$_{i+1}$ bonds. (In the graphic, CH represents a C$_\alpha$ atom and the attached hydrogen atom.) The torsion angle pair (ϕ, ψ) on either side of a residue R is considered a whole position. Three such pairs are shown. The torsion angle pair (ψ, ϕ) on either side of a peptide bond, between two residues, is considered a half position. Two such pairs are shown.

sity estimation methods. However, when the data are subdivided based on known characteristics such as amino acid residue or secondary structure type at the relevant sequence position, datasets quickly become small, sometimes having only a few dozen or a few hundred observations. A number of approaches to smooth density estimates from simple binning methods for the (ϕ, ψ) distributions have been proposed (Hovmoller et al., 2002; Lovell et al., 2003; Rother et al., 2008), but they behave poorly for these subdivided datasets. This is unfortunate, since these subsets provide structure predictions that are more accurate, as they utilizes more specific information about a particular sequence position. The issue is further complicated by the circular nature of this data, with each angle falling in the interval $(-\pi, \pi]$, which renders traditional techniques inadequate for describing the distributional characteristics. Distributions for angular data, particularly mixture distributions for bivariate angular data, are required.

Some methods have been proposed which exhibit better performance for small bivariate angular datasets. Pertsemlidis et al. (2005) recommend estimating such distributions using a finite number of Fourier basis functions. This method exhibits correct wrapping behavior, but requires the estimation of a large number of parameters which may not be readily interpretable. Other models exhibit more intuitive behavior. Mardia, Taylor, and Subramaniam (2007) fit finite mixtures of bivariate von Mises distributions using the EM algorithm. Dahl, Bohannan, Mo, Vannucci, and Tsai (2008) use a Dirichlet process mixture (DPM) model and bivariate normal distributions to estimate the distribution of torsion angles. However, neither of these methods is entirely satisfactory, as the first requires the selection of the number of component distributions, and the second cannot properly account for the wrapping of angular data.

We propose a nonparametric Bayesian model that takes the best aspects from Mardia et al. (2007) and Dahl et al. (2008). Specifically, we use a bivariate von Mises distribution as the centering and component distributions of a Dirichlet process mixture model. The use of a DPM model offers advantages in that the number of component distributions need not be fixed and inference accounts for the uncertainty in the number of components. Using a bivariate von Mises distribution, rather than a non-angular distribution, also provides estimates that properly account for the wrapped nature of angular data. In addition, the model readily permits the incorporation of prior information, which is often available for torsion angles.

Although some authors have studied Bayesian models for univariate angular data, to our knowledge the Bayesian analysis of bivariate angular data, such as that arising in protein structure prediction, has not been treated in the literature. We provide the results necessary for Bayesian analysis of bivariate angular data, including the full conditional distributions and conditionally conjugate priors, for a version of the bivariate von Mises distribution known as the sine model (Singh, Hnizdo, and Demchuk, 2002). Due to the complexity of this distribution, methods for sampling from the posterior distribution are not obvious. Therefore we provide an MCMC scheme that mixes well without requiring the tuning of any sampling parameters and show how to produce density estimates from the MCMC sampler.

We use our method to address the bioinformatics question of what distributions should be used when sampling to generate new candidate models for a protein's structure, a matter of considerable interest to the structure prediction community. Recall the illustration in Figure 1, which depicts whole and half positions on a peptide backbone. Current methods use data from whole positions, so the (ϕ, ψ) angle pairs across positions for an amino acid are considered independently. An alternative is to use the so-called half positions, which consist of ψ and ϕ angles on either side of a peptide bond. Treating data as half positions allows for more precise categorization, since these angle pairs are associated with two adjacent residues types, as opposed to a single residue for whole positions. Since they make use of a finer classification of the dataset, half position distributions are more accurate than those of the whole positions, thus providing a better description of backbone behavior. Due to their specificity, datasets for half positions are often relatively small, a situation that our proposed density estimation technique handles well.

Section 2.2 contains a review of past work in angular data analysis, including recent work in mixture modeling. In Section 2.3, we describe our DPM model for bivariate angular data that incorporates the von Mises sine model as a centering distribution in the Dirichlet process prior. In Section 2.4, we present the groundwork for a Bayesian treatment of the bivariate von Mises distribution and develop the relevant distribution theory, including deriving the full conditional distributions and conditionally conjugate priors for both the mean and precision parameters. We also describe our MCMC scheme for fitting this model, and our associated density estimation technique. Section 2.5 details the novel results from our method, comparing the use of whole versus half positions for template based protein structure modeling. Concluding comments are found in Section 2.6.

2.2 Review of Previous Statistical Work

As our method builds upon previous univariate and bivariate work with angular data, we provide a review of this field. We also discuss the recent results in bivariate mixture modeling. It should be noted that the terms *angular data* and *circular data* are used interchangeably in the literature.

2.2.1 Univariate Angular Data

A common option for describing univariate circular data is the von Mises distribution (see e.g. Mardia, 1975), which can be characterized in terms of either an angle or a unit vector. In terms of an angle $\phi \in (-\pi, \pi]$, the density is written:

$$f(\phi|\mu, \kappa) = \{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\phi - \mu)\}$$

where $\kappa > 0$ is a measure of concentration, μ is both the mode and circular mean, and $I_m(x)$ is the modified Bessel function of the first kind of order m. This distribution is symmetric and goes to a uniform distribution as $\kappa \to 0$. As discussed by Pewsey and Jones (2005), this distribution can be approximated by a wrapped normal distribution.

There is extensive Bayesian literature for this univariate distribution. Mardia and El-Atoum (1976) derived the full conditional distribution and conditionally conjugate prior for μ, while Guttorp and Lockhart (1988) determined the full conditional and conditionally conjugate prior for κ, as well as the conjugate prior and posterior distribution for simultaneous inference on μ and κ. Bagchi and Guttman (1988) developed the more general case including the distributions on the sphere and hypersphere. More recently, Rodrigues et al. (2000) presented an empirical Bayes approach to inference.

2.2.2 Bivariate Angular Data

The original bivariate von Mises distribution was introduced by Mardia (1975) and was defined with eight parameters. Rivest (1988) introduced a six parameter version. A five parameter distribution is preferable, however, so that the parameters might have a familiar interpretation, analogous to the bivariate normal.

Singh et al. (2002) introduced a five parameter subclass of Rivest's distribution referred to as the sine model. The density for angular observations (ϕ, ψ) is of the form:

$$f(\phi, \psi | \mu, \nu, \kappa_1, \kappa_2, \lambda) = C \exp\{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \lambda \sin(\phi - \mu) \sin(\psi - \nu)\} \tag{2.1}$$

for $\phi, \psi, \mu, \nu \in (-\pi, \pi]$, $\kappa_1, \kappa_2 > 0$, $\lambda \in (-\infty, \infty)$, and

$$C^{-1} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1 \kappa_2} \right)^m I_m(\kappa_1) I_m(\kappa_2). \tag{2.2}$$

This density is unimodal when $\lambda^2 < \kappa_1 \kappa_2$ and bimodal otherwise. In the unimodal situation, this density has a direct analogue to a bivariate normal with mean (μ, ν), and precision matrix Σ^{-1}, where $\Sigma^{-1}_{11} = \kappa_1$, $\Sigma^{-1}_{22} = \kappa_2$, and $\Sigma^{-1}_{12} = \Sigma^{-1}_{21} = -\lambda$. Note that this normal approximation holds when the variance of the distribution is small, i.e. when κ_1 and κ_2 are large. This correspondence to the bivariate normal distribution provides intuition for the behavior of the sine model for various parameter values.

Bivariate angular data, particularly protein conformational angles, often have a distribution with features that cannot be accommodated by a single von Mises distribution, even when bimodality is permitted. Mardia et al. (2007) developed the cosine model, another five parameter bivariate angular distribution, and suggested using the EM algorithm to fit several finite mixtures of these models, each with a different numbers of components. They employed the Akaike information criterion (AIC) for model selection. With this technique they estimated the density of (ϕ, ψ) angle pairs in the myoglobin and malate dehydrogenase protein structures.

2.3 Bayesian Mixture Model with von Mises Distributions

Our model for bivariate angular distributions offers both the flexibility of the DPM model and the technical accuracy provided by the use of a bivariate angular distribution. The proposed model is:

$$\begin{aligned}(\phi_i, \psi_i) \mid \mu_i, \nu_i, \Omega_i &\sim p((\phi_i, \psi_i) \mid \mu_i, \nu_i, \Omega_i) && (2.3)\\ (\mu_i, \nu_i, \Omega_i) \mid G &\sim G && (2.4)\\ G &\sim DP(\tau_0 H_1 H_2), && (2.5)\end{aligned}$$

where $p((\phi_i, \psi_i)|\mu_i, \nu_i, \Omega_i)$ is a bivariate von Mises sine model in which Ω_i is a 2×2 matrix with both off-diagonal elements equal to $-\lambda_i$ and diagonal elements κ_{1i} and κ_{2i}. This parameterization makes Ω_i analogous to the precision matrix of the bivariate normal distribution. The distribution G is a random realization from $DP(\tau_0 H_1 H_2)$, a Dirichlet process (Ferguson, 1973) with mass parameter τ_0 and centering distribution $H_1 H_2$. We take H_1 to be a bivariate von Mises sine model for the means μ and ν, and H_2 to be a bivariate Wishart distribution for the precision matrix Ω. An alternative noninformative prior on the means is obtained using a uniform distribution on the square $(-\pi, \pi] \times (-\pi, \pi]$ for H_1. In either case, the result is a Bayesian mixture model (Antoniak, 1974), a broad class of models reviewed by Müller and Quintana (2004).

In contrast, Dahl et al. (2008) modeled the distributions of conformational angles using a DPM model that assumed bivariate normals as the component distributions. They took the sampling model to be a bivariate normal distribution with precision matrix Σ_i^{-1} and also set H_1 to be a bivariate normal. This approach is unsatisfactory for circular data and exhibits particular problems when the underlying distribution has significant mass on the boundaries of the $(-\pi, \pi] \times (-\pi, \pi]$ region. Our use of the bivariate von Mises distribution avoids this deficiency. Also, in contrast to our model, Dahl et al. (2008) used two separate clusterings: one for the mean parameters and one for the precision parameters.

For our torsion angle application, we are particularly interested in predicting new (ϕ, ψ) values based on the existing data and our DPM model. Density estimation using DPM models is discussed by Escobar and West (1995). A nonparametric density estimate of the (ϕ, ψ) space from data $(\boldsymbol{\phi}, \boldsymbol{\psi}) = ((\phi_1, \psi_1), ..., (\phi_n, \psi_n))$ is a predictive distribution for a new angle pair (ϕ_{n+1}, ψ_{n+1}), namely:

$$\begin{aligned}
p((\phi_{n+1}, \psi_{n+1})|(\boldsymbol{\phi}, \boldsymbol{\psi})) &= \int p((\phi_{n+1}, \psi_{n+1}), (\mu_{n+1}, \nu_{n+1}, \Omega_{n+1})|(\boldsymbol{\phi}, \boldsymbol{\psi})) \\
&\quad d(\mu_{m+1}, \nu_{n+1}, \Omega_{n+1}) \\
&= \int p((\phi_{n+1}, \psi_{n+1})|(\mu_{n+1}, \nu_{n+1}, \Omega_{n+1})) \\
&\quad \times p((\mu_{n+1}, \nu_{n+1}, \Omega_{n+1})|(\boldsymbol{\phi}, \boldsymbol{\psi}))d(\mu_{n+1}, \nu_{n+1}, \Omega_{n+1}). \quad (2.6)
\end{aligned}$$

We show in the next sections how to estimate this density and how it can be used for protein structure prediction.

2.4 Model Estimation

The integral of the posterior predictive density in (2.6) cannot be expressed in closed form, but it can be computed through Monte Carlo integration. Specifically, let $(\mu_{n+1}^1, \nu_{n+1}^1, \Omega_{n+1}^1), \ldots, (\mu_{n+1}^B, \nu_{n+1}^B, \Omega_{n+1}^B)$ be B samples from the posterior predictive distribution of $(\mu_{n+1}, \nu_{n+1}, \Omega_{n+1})$ obtained from some valid sampling scheme. Then

$$p((\phi_{n+1}, \psi_{n+1})|(\boldsymbol{\phi}, \boldsymbol{\psi})) \approx \frac{1}{B} \sum_{b=1}^{B} p((\phi_{n+1}, \psi_{n+1})|(\mu_{n+1}^b, \nu_{n+1}^b, \Omega_{n+1}^b)). \quad (2.7)$$

While equation (2.7) can be evaluated for any value of (ϕ_{n+1}, ψ_{n+1}), for our purposes we obtain density estimates by evaluating (2.7) on a grid of points and use linear interpolation between them.

All that remains is to determine how to sample from the posterior distribution of the parameters. The Auxiliary Gibbs sampler of Neal (2000) provides an MCMC update of the allocation of observations to clusters. We are at liberty to choose any valid updating scheme for the mean and precision parameters. Since the joint posterior distribution for all five parameters is intractable, the full conditionals of the mean and precision parameters are a natural choice. We now present our novel results regarding: 1) conditionally conjugate priors for this model, 2) full conditional distributions for both conditionally conjugate and uniform priors, and 3) approximate sampling methods for each full conditional distribution.

2.4.1 Full Conditional Distributions of Mean and Precision Parameters

Using the notation from Mardia et al. (2007), the eight parameter bivariate von Mises distribution may be expressed as:

$$\begin{aligned} f(\phi,\psi) \propto \exp\{&\kappa_1 \cos(\phi-\mu) + \kappa_2 \cos(\psi-\nu) + \\ &[\cos(\phi-\mu), \sin(\phi-\mu)] A [\cos(\psi-\nu), \sin(\psi-\nu)]^T\} \end{aligned}$$

where A is a 2×2 matrix of association parameters. The sine model density from (2.1) corresponds to the situation in which $A_{11} = A_{12} = A_{21} = 0$ and $A_{22} = \lambda$.

The conditionally conjugate prior for the mean parameters, whether observations are from an eight parameter or sine model bivariate von Mises distribution, is an eight parameter bivariate von Mises distribution. We are particularly interested in using a sine model prior with center (μ_0, ν_0) and precision parameters κ_{10}, κ_{20}, and λ_0. This prior can be interpreted as an additional observation with known precision parameters. As observations with higher concentration values have greater weight in determining the posterior distribution parameters, less informative priors are those with κ_{10}, κ_{20}, and λ_0 close to 0. This is consistent with the fact that an alternative noninformative prior is a uniform distribution on $(-\pi, \pi] \times (-\pi, \pi]$, which is the limit of the sine model prior when $\lambda_0 = 0$ and $\kappa_{10}, \kappa_{20} \rightarrow 0$.

Consider a set of observations (ϕ_i, ψ_i), $i = 1, ..., n$, each with known precision parameters κ_{1i}, κ_{2i}, and λ_i. The full conditional distribution for (μ, ν) is an eight parameter bivariate von Mises distribution with full derivation and details given in Appendix A. The full conditional parameters are:

$$\tilde{\mu} = \arctan\left(\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right) \qquad \tilde{\nu} = \arctan\left(\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right)$$

$$\tilde{\kappa}_1 = \left|\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right| \qquad \tilde{\kappa}_2 = \left|\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right|$$

$$\tilde{A} = \sum_{i=0}^{n} \lambda_i \left[\begin{array}{cc} \sin(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & -\sin(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \\ -\cos(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & \cos(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \end{array}\right]. \qquad (2.8)$$

The mean parameters of the full conditional distribution are the directions of the sums of the observation vectors, while the concentration parameters are the magnitudes of those same vectors. These bivariate results are analogous to the univariate work of Mardia and El-Atoum (1976).

When considering the full conditional distribution of the precision parameters of the sine model, it may be assumed that the known means are both 0. The conditionally conjugate prior for the precision parameters is of the form:

$$\pi(\kappa_1, \kappa_2, \lambda) \propto \left\{4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_m(\kappa_2)\right\}^{-c}$$
$$\times \exp(R_{\phi 0}\kappa_1 + R_{\psi 0}\kappa_2 + R_{\phi\psi 0}\lambda). \qquad (2.9)$$

Here the prior assumes the role of c observations from the bivariate von Mises sine model, and the prior parameters $R_{\phi 0}$ and $R_{\psi 0}$ are the sums of the magnitudes in the x direction of the ϕ and ψ components, respectively, of these observations. The parameter $R_{\phi\psi 0}$ is the sum of the products of the magnitudes in the y direction. For this interpretation to hold, $R_{\phi 0}$, $R_{\psi 0}$, and $R_{\phi\psi 0}$ must be between $-c$ and c. Notice that the conditionally conjugate prior, and corresponding full conditional distribution, are difficult to sample from due to the infinite sums of Bessel functions. Notice also that this prior does not guarantee precision parameters that will give unimodal sine model distributions.

2.4.2 Markov Chain Monte Carlo Sampler

The posterior distribution of our model parameters from Section 2.3 can be sampled through Markov chain Monte Carlo using the Auxiliary Gibbs sampler of Neal (2000). This method requires the ability to directly sample from the centering distribution. For this reason we use a bivariate von Mises sine model, rather than an eight parameter bivariate von Mises distribution, for H_1. It is also difficult to sample from the conjugate prior for the precision parameters described in (2.9), and we instead use the Wishart distribution for $H_2(\Omega)$ in (2.5). In addition, a Wishart prior guarantees that the sampled matrix will be positive definite, which is equivalent to the restriction that ensures unimodality for the sine model component distributions. Eliminating bimodality both simplifies posterior simulation, and increases the resemblance of the sampling model to that of a mixture of bivariate normal distributions. This substitution is also appealing because, for large values of κ_1 and κ_2, this von Mises model is nearly equivalent to a normal distribution. In this case, the Wishart prior behaves much like the conjugate prior distribution in (2.9).

Auxiliary Gibbs sampling requires a valid updating scheme for the model parameters. Generating MCMC samples for the full conditional distribution of the means is fairly straightforward. As it is difficult to sample directly from the eight parameter full conditional distribution, we instead generate proposals using a sine model as part of an independence sampler. The parameters of our proposal distribution are:

$$\tilde{\mu} = \arctan\left(\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right) \qquad \tilde{\nu} = \arctan\left(\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right)$$

$$\tilde{\kappa}_1 = \left|\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right| \qquad \tilde{\kappa}_2 = \left|\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right|.$$

$$\tilde{\lambda} = \left(\sum_{i=0}^{n} \lambda_i \cos(\phi_i - \psi_i) \right) \{\cos(\tilde{\mu} - \tilde{\nu})\}^{-1} .$$

This distribution uses the mean and concentration parameters from the true full conditional distribution, altering only the parameters used to model association. The chosen dependence parameter $\tilde{\lambda}$ has been found to work well in practice.

A simple method to sample from the sine model is to use a rejection sampler with a uniform distribution as the majorizing density. The implementation requires some care, however, as the full conditional distribution is not always unimodal. The value of the mode in the unimodal case is $(\tilde{\mu}, \tilde{\nu})$, while the values in the bimodal case depend on the sign of $\tilde{\lambda}$ and are given in the appendix of Mardia et al. (2007).

The update scheme for the concentration matrix Ω of a cluster is less straightforward. Regardless of the choice of prior, the full conditional distribution of the precision parameters would be difficult to sample from directly, due to the infinite sum of Bessel functions and the fact that the constant of integration is not known in closed form. However, this distribution is often well approximated by the full conditional of the precision parameters from an analogous model in which the data are assumed to be normally distributed, particularly when a Wishart prior is used. An independence sampler using this equivalent Wishart distribution generally provides a good acceptance rate. Further, this proposal distribution is automatic in the sense that the resulting sampling scheme does not require any tuning parameters. The use of this proposal distribution is also consistent with previous findings for the univariate case, where the full conditional distribution of κ was found to be approximately χ^2 distributed (Bagchi and Guttman, 1988).

An outline of this algorithm is given in Figure 2.

1. Initialize the parameter values:
   (a) Choose an initial clustering. Two obvious choices are: (1) one cluster for all of the angle pairs, or (2) each angle pair in a cluster by itself.
   (b) For each initial cluster S of observed angle pairs, initialize the value of the common bivariate von Mises parameters μ, ν, Ω by sampling from the centering distribution $H_1(\mu, \nu)H_2(\Omega)$ of the DP prior.
2. Obtain draws from the posterior distribution by repeating the following:
   (a) Given the mean and precision values, update the clustering configuration using one scan of the Auxiliary Gibbs sampler of Neal (2000).
   (b) Given the clustering configuration and precision values, update the values of (μ, ν) for each cluster using the independence sampler in Section 2.4.2.
   (c) Given the clustering configuration and mean values, update the precision matrix Ω for each cluster using the Wishart independence sampler described in Section 2.4.2.

Figure 2: Summary of computational procedure for density estimation with angle pairs.

2.5 Template Based Modeling of Protein Structure

2.5.1 Motivation

In this section we use our proposed density estimation procedure to develop a more efficient method for protein structure prediction. Methods specifically designed for angular data are necessary since consideration of periodicity is essential for certain amino acids, such as glycine. Figure 3 shows density estimates based on the normal model of Dahl et al. (2008) and our own von Mises sine model. Notice that the normal model is unable to wrap between the angles $-\pi$ and π. The von Mises model identifies a single peak that includes mass at all four corners, whereas the normal model identifies separate peaks at each corner for this same portion of the data.

We also conducted a quantitative comparison of these two DPM models. To investigate the improvement of the von Mises over the normal, we generated density estimates for subsets of size 200 for each of the twenty amino acid datasets, once using normal cen-

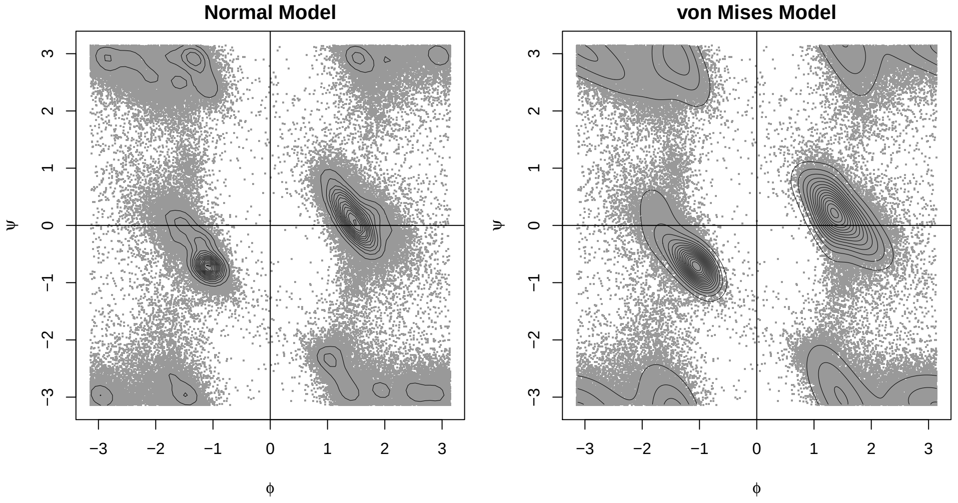


Figure 3: Ramachandran plots for the 121,497 angle pairs that make up the PDB dataset for the residue glycine, along with density estimates based on both the normal and von Mises distributions. The normal model is from the work of Dahl et al. (2008) while the von Mises estimate is based on our model in Section 2.3. Note that glycine spans almost the complete range of values in both ϕ and ψ, which makes the use of a method that correctly models circular data critical.

tering and component distributions and once using the equivalent von Mises distributions. In each case we used the prior parameter settings and clustering configuration from Dahl et al. (2008), with separate clusterings for mean and precision parameters. We calculated the Bayes factor for the two models using the full amino acid datasets, which ranged in size from 23,000 to 143,000 observations. Our Bayes factor was defined as:

$$B((\boldsymbol{\phi},\boldsymbol{\psi})) = \frac{p((\boldsymbol{\phi},\boldsymbol{\psi})|M_1)}{p((\boldsymbol{\phi},\boldsymbol{\psi})|M_2)}$$

where M_1 was our von Mises model estimate and M_2 was normal model estimate. The logs of the Bayes factors ranged from 183 to 6,013 in absolute value, allowing us to draw clear conclusions as to the superior model in each case. For nineteen of the twenty amino acids, the Bayes factor indicated that the von Mises model was superior. While the normal model fails to capture the wrapped nature of torsion angle data, our method provides robust and elegant estimates of the (ϕ, ψ) distributions from large or small datasets.

We can use our nonparametric density estimation procedure to estimate the density of backbone torsion angle distributions. This approach allows us to investigate how well distributions obtained from Protein Data Bank (PDB) data approximate the (ϕ, ψ) distributions at particular positions in a protein fold "family". This is of interest since one popular technique in protein structure prediction is to generate candidate conformations based on the structures of known similar proteins. These fold families can provide a great deal of information about the unknown structure, but most are very small, often with fewer than 10 members. This means that density estimation purely within a family has not been feasible. In such cases, candidate distributions are generated based on large datasets with similar characteristics to those of the sequence positions in the known structures. As current search methods are mostly random walks in conformation space (Dill et al., 2007; Lee and Skolnick, 2008; Das and Baker, 2008), improved modeling of these positional densities increases the chance of finding a good structure. To assess the quality of these PDB "category densities," we compare density estimates from the PDB to those obtained from three fold families: globins, immunoglobulins, and TIM barrels. Each represents a classic architecture in structural biology. The globins consist mostly of α-helical secondary structure, and the immunoglobulins consist mostly of β-sheets. TIM barrels are a mixed structure with both α-helices and β-sheets. These three families are fairly unique in that they have enough known members that density estimation purely within a family is possible.

In contrast to standard methods, we not only consider the torsion angles around a sequence position or residue, but also the (ψ, ϕ) torsion angle pair around the peptide bond (see Figure 1). Previously, this peptide centered view of torsion angles has only been applied to short amino acid chains (Anderson and Hermans, 1988; Grail and Payne, 2000). Recall that we refer to the residue torsion angle pairs (ϕ, ψ) as "whole positions" and the peptide torsion angle pairs (ψ, ϕ) as "half positions," since they reside "half-way" between whole sequence positions. By incorporating the characteristics of two residues, these half positions lead to a finer classification of the dataset, and provide an effective approach to increasing the amount of information known about a particular angle pair without increasing the complexity of the underlying model beyond two torsion angles.

Each whole position can be described by which of twenty amino acid residues is present, and also the type of secondary structure at that location. We define secondary structure in the same manner as the Definition of Secondary Structure for Proteins (DSSP) program (Kabsch and Sander, 1983). The normal eight classes are condensed to four: helices (H), sheets (E), coils (C), and turns (T). Residues without any specific structure are assigned to the random coil (C) class. β-turns and G-turns were combined into the turn (T) class. All helices were classified as (H). Strand and β-bulges were combined into the extended strand (E) class. The twenty residues and four secondary structure classes provide eighty possible classifications for whole position data.

Since a half position involves two residues, there are 400 categories when considering only amino acid pairs, and 6,400 when the four secondary structure classes are included. When considering half positions, we take the same data as used for the whole positions and divide it into a much larger number of groups, which thins out the data considerably. This reduction is worthwhile, however, since every amino acid and secondary structure type exhibits unique behavior visible on the Ramachandran plot. Using adjacent pairs of amino acids and structure types, as the half positions do, gives even more specific information about a sequence position. As we will demonstrate, the use of half positions provides a substantial increase over the available information provided by whole position data.

2.5.2 Methods and Diagnostics

The torsion angle distributions were estimated for the PDB whole and half positions, as well as the three families of protein folds: globins, immunoglobulins, and TIM barrels. For whole positions, in addition to the categories discussed before, we include a category ignoring secondary structure type for a total of 100 density estimates. The same was done for half position densities, giving a total of 6,800 estimates.

For each of the three protein fold families, angle pairs for whole and half positions were obtained for each sequence position. For instance, all 92 (ϕ, ψ) pairs at position 13 based on the globin alignment were used to estimate the relevant density. The same was done for half positions, but the (ψ, ϕ) angles were centered around the peptide bond between two residues. These alignments produced 183 residue positions for the globins, 343 for the immunoglobulins, and 274 for the TIM barrels.

For each dataset, two chains were run for 6,000 iterations, with the first 1,000 discarded as burn in. For post burn in iterations, a draw was taken from the posterior distribution and the resulting von Mises density was evaluated for a grid of 360×360 points. Using 1 in 10 thinning, this gave $B = 1,000$ samples to estimate the density using (2.7). For datasets with over 2,000 observations, we used a random subsample of 2,000 observations.

Our von Mises model from Section 2.3 was used with mean prior parameters $\mu_0 = \nu_0 = 0$, and Ω_0 was a diagonal matrix with elements $1/\pi^2$. The small concentration values made this prior largely noninformative. For the Wishart prior, we used used $v = 2$ degrees of freedom and set the scale matrix B to have diagonal elements of 0.5^2, and off-diagonal elements of 0 (making the expected value $\frac{v}{2}B^{-1} = B^{-1}$). This again provided a diffuse centering distribution on the radian scale. The mass parameter τ_0 of the Dirichlet process was set to 1.

Convergence was evaluated using entropy as described by Green and Richardson (2001). Figure 4 shows trace plots for the two MCMC chains for position 11 of the globin family.

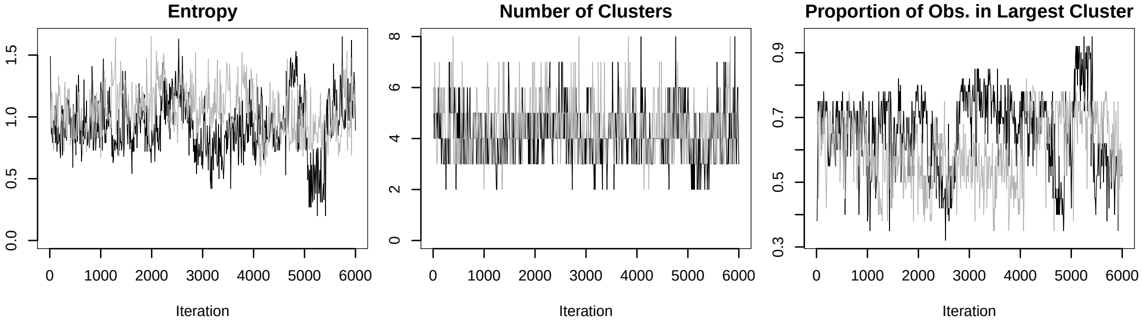


Figure 4: Convergence diagnostics for globin position 11.

2.5.3 Comparison of Whole and Half Position Density Estimates

To judge whether the whole or half position density estimates provided a closer match to the density at a particular position of a protein family, we used the Jensen-Shannon divergence:

$$\frac{1}{2}\left(D_{KL}(P, \frac{P+Q}{2}) + D_{KL}(Q, \frac{P+Q}{2})\right)$$

as a measure of distributional similarity, where D_{KL} is the Kullback-Leibler divergence defined by $D_{KL}(P,Q) = \sum_i P(i)\log\left(P(i)/Q(i)\right)$. Both P and Q are density estimates from our proposed procedure.

The positional density estimates were compared to all of the estimates from the PDB using this divergence score. Whole position densities from each of the three fold families were compared to the whole position category densities from the PDB, and half positions from the fold families were compared to the half position category densities from the PDB. The best matches, those with the lowest divergence values, are plotted against position in Figure 5. It is evident that the half position comparisons produce lower divergence scores. The mean minimum divergence for whole positions is 0.143, while the corresponding half position value is 0.052. The paired sign test of the null hypothesis that the median minimum divergence score for whole positions is less than or equal to that for half positions produced p-values less than 0.0001 for each structure family. The plot shows that the half

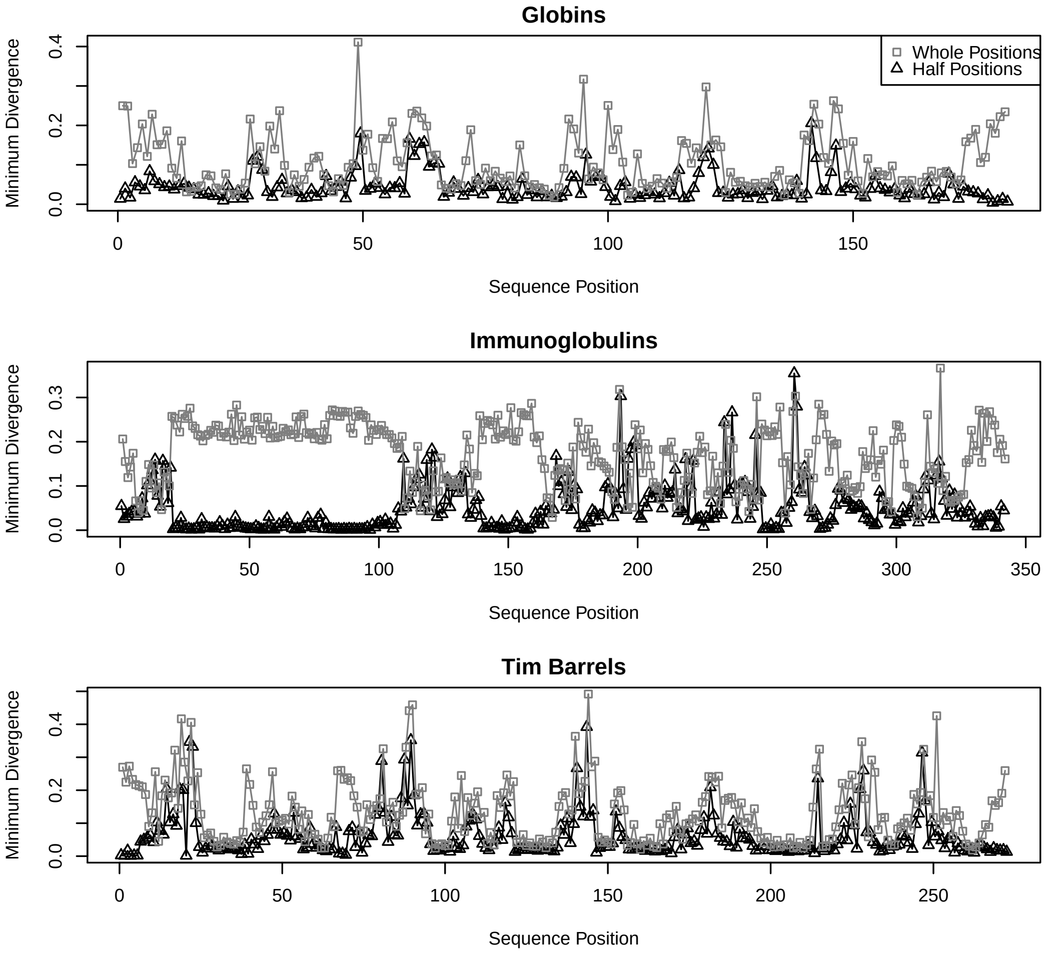


Figure 5: A comparison of minimum divergence scores for whole versus half positions.

positions provide better matches at the beginning and ends of the structures, which consist of coil secondary structure, and in the sheet regions of the immunoglobulins. Whole positions perform best in helical regions, but even then half positions provide a better match. The worst matching cases are in areas with non-canonical turns or unique coils, which correspond to the highest minimum divergence scores for all structure families.

A specific example of this behavior can be seen in Figure 6, which shows the globin whole position 11 with the closest matching PDB density compared to the half position 11-12 with its matching half position density from the PDB. It can be readily seen from these

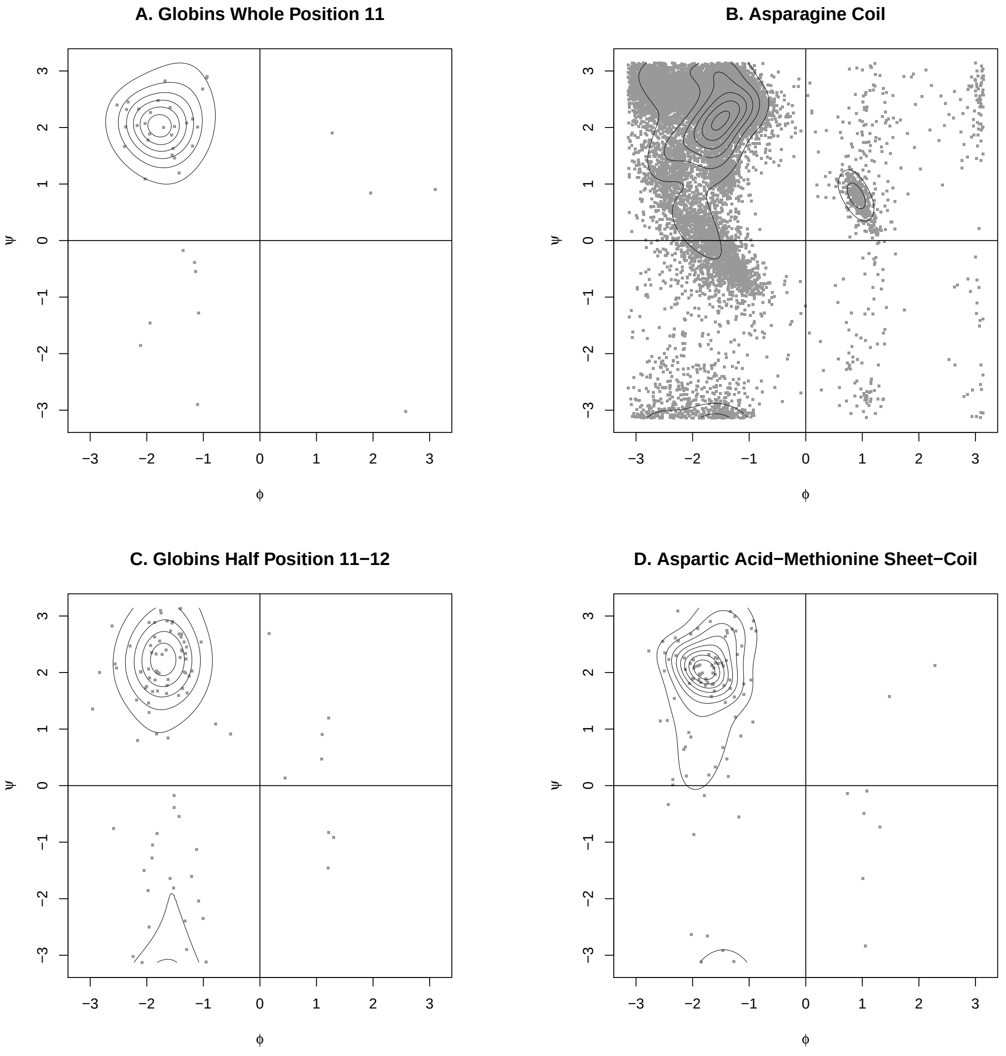


Figure 6: Ramachandran plots around globin position 11. A) Density estimate and data for whole position 11. B) Density estimate and data for asparagine coil whole positions which, at a divergence of 0.092, provides the best PDB match for globin whole position 11. C) Density estimate and data for the half position between residues 11 and 12. D) Density estimate and data for aspartic acid sheet to methionine coil half positions which, at a divergence of 0.037, provides the best PDB match for globin half position 11-12.

figures that the whole position matches fairly well, but also includes extraneous density. By instead considering the half position of the associated peptide, we find a closer match. This is not surprising due to the increased specificity of the half position densities from the PDB, not to mention the increased number of categories available for comparison. These results suggest that the use of half position data as a substitute for whole position data provides better results.

2.6 Discussion

We have presented a novel nonparametric Bayesian method for density estimation with bivariate angular data. This method, unlike many currently used to estimate the density of (ϕ, ψ) angle pairs, provides smooth estimates without requiring large datasets. This allowed the estimation of the distributions for PDB half position data, as well as positional data from three protein fold families. Using this new technique we were able to evaluate the common practice of using whole position estimates for positional data. Our results indicate that half position densities are more informative than the corresponding whole position estimates.

Our Dirichlet process mixture model performs well for density estimation of bivariate circular data. In contrast to previous work in this area, it does not require the setting of a fixed number of components for the mixture. By incorporating the bivariate von Mises sine model, we are able to account for the wrapping of the data, and the sine model's equivalence to the normal distribution allows for a straightforward interpretation and effective implementation of a Markov chain Monte Carlo sampling scheme. This was made possible by our results regarding the full conditional distributions for the mean and precision parameters.

We have demonstrated that our approach at half positions provides greater precision than the use of whole positions for protein structure prediction. Unlike the fold families

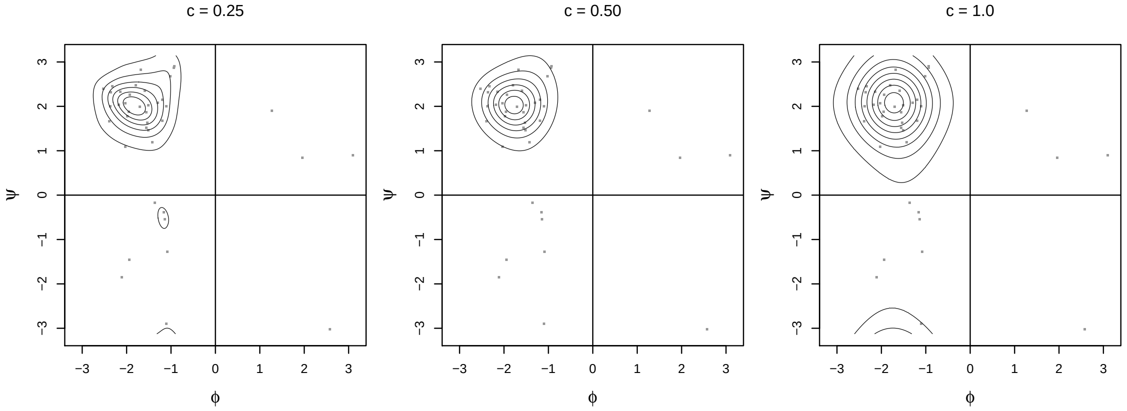


Figure 7: Density estimates for globin position 11 with different scale matrices for the Wishart prior distribution.

shown here, most protein folds have very limited numbers of representatives in the PDB. For these fold families, density estimation at each position, even using our method, is not feasible. Therefore, the distributions used to approximate the backbone torsion angle space are obtained from the PDB. When these distributions are inaccurate or too broad, as we see for the whole positions, significant time is spent sampling the wrong areas of backbone conformation space. When searching using a random walk in conformation space, this reduces the chance of finding a good structure. A reliable reduction of the backbone search space using the half position distributions is a significant improvement to all structure prediction methods. The only way such half position distributions can be precisely calculated is by using density estimation methods, such as ours, that properly address the angular nature of the data and cope well with smaller datasets.

We conclude by briefly presenting the results of a sensitivity analysis we performed for the Wishart prior and DP mass parameter. Three different scale matrices were considered for the Wishart prior. Each could be written as $c^2 I$, where I was the 2×2 identity matrix, and c took values 0.25, 0.5, and 1.0. Figure 7 shows the resulting density estimates for globin position 11. The changes between the density estimates are not dramatic, and

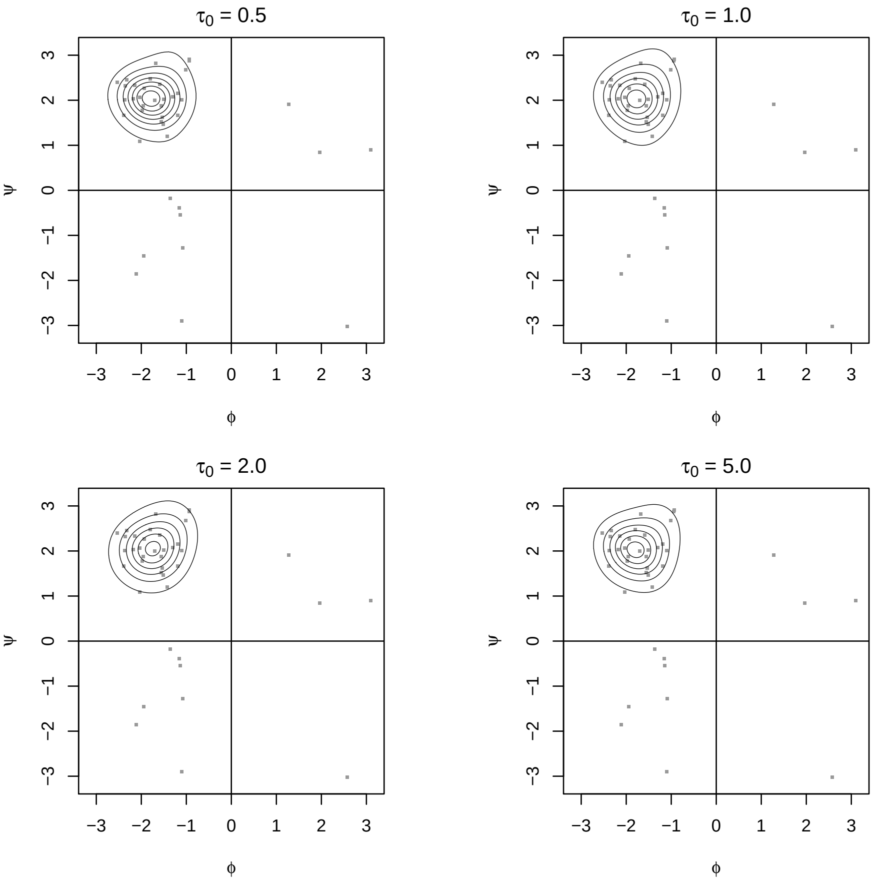


Figure 8: Density estimates for globin position 11 for assorted values of the mass parameter.

the effect is comparable to that of varying the bandwidth in kernel density estimation methods. Other positions showed similar behavior, although the effect of changing the prior parameter is reduced as sample size increases..

We also investigated the sensitivity to changes in the mass parameter. We set τ_0 to 0.5, 1.0, 2.0, and 5.0. A comparison of these estimates for position 11 is given in Figure 8. The plots all look very similar. This is generally the behavior of the other positions, although sometimes the 5.0 case exhibits slight but noticeable differences.

Convergence of the Markov chains was generally good, but we did encounter occasional difficulties, particularly when the mass parameter was small. However, even when the trace plots of entropy for the two chains suggested convergence problems, the density estimates generated by the separate chains were generally similar. Therefore, we do not consider this to be a major issue. On the other hand, if the mass parameter is very small severe problems can occur. As always, convergence diagnostics should be employed.

CHAPTER III

A DIRICHLET PROCESS MIXTURE OF HIDDEN MARKOV MODELS FOR PROTEIN STRUCTURE PREDICTION*

3.1 Introduction

The field of protein structure prediction has greatly benefited from formal statistical modeling of available data (Osguthorpe, 2000; Bonneau and Baker, 2001). More automatic methods for predicting protein structure are critical in the biological sciences as they help to overcome a major bottleneck in effectively interpreting and using the vast amount of genomic information: determining the structure, and therefore the function, of a gene's protein product. Currently the growth of genomic data far outstrips the rate at which experimental methods can solve protein structures. To help accelerate the process, protein structure prediction methods aim to construct accurate three-dimensional models of a target protein's native state using only the protein's amino acid sequence.

Protein structure is typically described in terms of four categories: primary through quaternary. Primary structure consists of the linear sequence of covalently bonded amino acids that make up a protein's polypeptide chain. Secondary structure describes the regularly repeating local motifs of α-helices, β-strands, turns, and coil regions. For a single polypeptide chain, tertiary structure describes how the secondary structure elements arrange in three-dimensional space to define a protein's fold. By allowing the polypeptide chain to come back on itself, the loops and turns effectively define the arrangement

* Material in this chapter is reprinted with permission from "A Dirichlet Process Mixture of Hidden Markov Models for Protein Structure Prediction" by Lennox, K. P., Dahl, D. B., Vannucci, M., Day, R. and Tsai J. 2010. *The Annals of Applied Statistics*. In press.
Journal URL: http://www.imstat.org/aoas/

of the more regular secondary structure of α-helices and β-strands. Quaternary structure describes how multiple folded polypeptide chains interact with one another. In a typical structure prediction problem the primary structure is known, and the goal is to use this information to predict the tertiary structure.

One of the standard approaches to this problem is template-based modeling. Template-based methods are appropriate when the target sequence is similar to the sequence of one or more proteins with known structure, essentially forming a protein fold "family." Typically the core of the modeled fold is well defined by regular secondary structure elements. One of the major problems is modeling the loops and turns: those regions that allow the protein's tertiary structure to circle back on itself. Unlike the consistency of the core in a template-based prediction, the variation in the loops and turns (both in terms of length and amino acid composition) between structures with the same fold family is often quite large. For this reason current knowledge-based methods do not use fold family data. Instead of the template-based approach, they use libraries of loops which are similar in terms of length and amino acid sequence to the target. However, such library datasets do not have the same level of structural similarity as do purely within-family datasets. In this work, our approach to modeling structural data allows us to effectively extend template-based modeling to the loop and turn regions and thereby make more informed predictions of protein structure.

Our approach is based on the simplest representation of protein structure: the so-called backbone torsion angles. This representation consists of a (ϕ, ψ) angle pair at each sequence position in a protein, and it provides a reduction in complexity from using the 12 Cartesian coordinates for the four heavy backbone atoms at each position. This method for describing protein structure was originally proposed by Ramachandran et al. (1963), and the customary graphical representation of this type of data is the Ramachandran plot. The Ramachandran plot in Figure 9 shows the (ϕ, ψ) angles of protein positions containing the amino acid alanine. The pictured dataset was obtained from the Protein Data Bank (PDB,

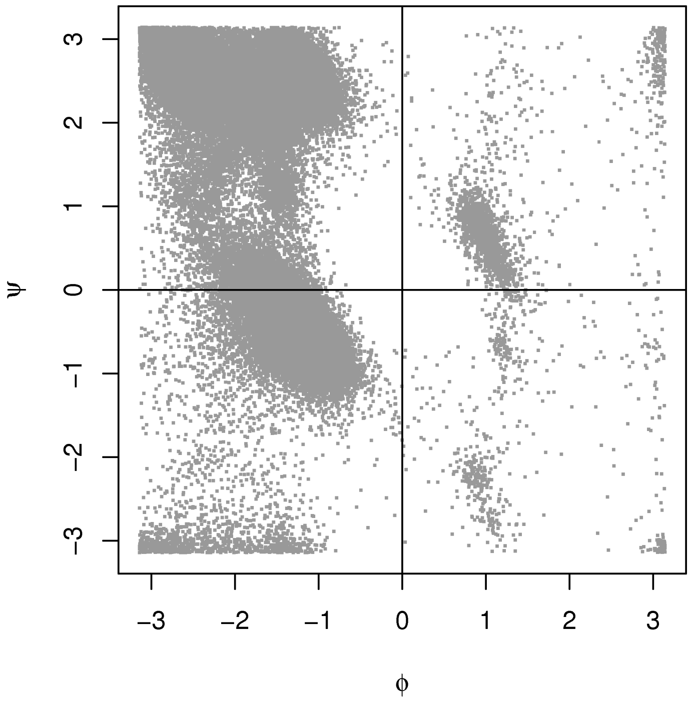


Figure 9: Ramachandran plot for the 130,965 angle pairs that make up the PDB dataset for the amino acid alanine. Angles are measured in radians.

Kouranov et al., 2006), a repository of solved protein structures.

Density estimation of Ramachandran space is particularly useful for template-based structure prediction. Because a target protein with unknown tertiary structure is known to be related to several proteins with solved structures, models for bivariate angular data can be used to estimate the distribution of (ϕ, ψ) angles for a protein family, and thereby generate candidate structures for the target protein.

While there has been considerable recent work on modeling in Ramachandran space at a single sequence position (see e.g. Ho et al., 2003; Lovell et al., 2003; Butterfoss et al., 2005), models that accommodate multiple sequence positions remain uncommon. A notable exception is the DBN-torus method of Boomsma et al. (2008). However this approach was developed primarily to address sampling of fragments in *de novo* protein structure prediction, and so specifically does not include protein family information. *De novo* structure prediction is used when similar proteins with known structure are unavailable and is thus inherently more difficult and less accurate than template based modeling. While template-based methods can draw on a certain amount of known information, a common complication is that protein families typically have fewer than 100 members, and often fewer than 30 members.

Not only do protein families tend to have few members, but the data within a family is "sparse," particularly in loop regions. A template sequence for a protein structure family is generated by simultaneously aligning all of the member proteins using amino acid type at each sequence position. The sequences in a fold family are often of different lengths due to different sizes of loops and turns. In such an alignment a typical member protein is not represented at every sequence position. This leads to what we call a "sparse data" problem. Note that this is not a missing data situation, as a sequence position is not merely unobserved, but rather does not in fact exist.

A joint model for a large number of torsion angles using somewhat limited data can be enhanced by leveraging prior knowledge about the underlying structure of the data. We present a Bayesian nonparametric model incorporating a Dirichlet process (DP) with one of two possible families of centering distributions for modeling the joint distributions of multiple angle pairs in a protein backbone. Our model addresses the sparse data situation, and also accommodates a larger number of sequence positions than previously considered methods of template-based density estimation. One of our proposed centering distributions leads to a largely noninformative prior, but we also propose a family of centering distributions based on known characteristics of protein secondary structure in the form of a hidden Markov model (HMM). The inclusion of an HMM allows our model to share structural information across sequence positions. Since each secondary structure type has a distinctive footprint on the Ramachandran plot, with this process we can use an informative prior to incorporate additional information into our model.

There is precedent for the use of a hidden Markov model for protein structure prediction in the DBN-torus model of Boomsma et al. (2008). There, secondary structure information is incorporated into the state space of a dynamic Bayesian network, a generalization of an HMM, which allows the DBN-torus model to infer secondary structure when generating candidate angle pair sequences. The model generates significantly better candidates, however, when secondary structure is provided from an external prediction method. There are other differences between the DBN-torus method and our own which result from the distinct applications of the two methods. DBN-torus is used for *de novo* structure prediction; it is designed to make predictions for any kind of protein, and is not customized for a particular fold family. In contrast, our method is tailored for template-based modeling. Thus, the DBN-torus model can be used even when template information is unavailable, but will miss opportunities for improvement when fold-family structure information exists.

We apply our method to the loop region between the E and F α-helices of the globin protein template, which varies between 8 and 14 sequence positions in length. By borrowing strength from neighbors containing numerous observations, our model generates informative density estimates even if relatively little data is available at a given position. This property gives our method a significant advantage in loop prediction by allowing the use of fold family data. This extension of template-based modeling to loop regions was not possible before the development of these statistical tools. We show that using our Dirichlet process mixture of hidden Markov models (DPM-HMM) in a template-based approach provides a better match to real structure data than does either a library-based method or DBN-torus.

In Section 3.2 we give some background on previous work in torsion angle modeling, as well as the bivariate von Mises distribution and the Dirichlet process. In Section 3.3 we present our model along with the informative and noninformative priors. An explanation of how to fit this model and use it for density estimation is provided in Section 3.4. Section 3.5 contains an application of our method to estimate the joint density of torsion angles in the EF loop region in the globin protein family. Finally, we discuss our conclusions in Section 3.6.

3.2 Preliminaries

We illustrate the development of our model by first discussing methods for modeling individual torsion angle pairs. Working with torsion angles requires the use of distributions specifically designed to account for the behavior of angular data. This data has the property that an angle ϕ is identical to the angle $\phi + 2k\pi$ for all $k \in \{..., -1, 0, 1, ...\}$. The bivariate von Mises distribution is commonly used for paired angular data.

Originally proposed as an eight parameter distribution by Mardia (1975), subclasses of the bivariate von Mises with fewer parameters are considered easier to work with and are often more interpretable. Rivest (1982) proposed a six parameter version, which has been further refined into five parameter distributions. One such subclass, known as the cosine model, was proposed by Mardia et al. (2007), who employed it in frequentist mixture modeling of (ϕ, ψ) angles at individual sequence positions. We consider an alternative developed by Singh et al. (2002) known as the sine model.

The sine model density for bivariate angular observations (ϕ, ψ) is defined as:

$$f(\phi, \psi|\mu, \nu, \kappa_1, \kappa_2, \lambda) = C \exp\{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \lambda \sin(\phi - \mu) \sin(\psi - \nu)\} \tag{3.1}$$

for $\phi, \psi, \mu, \nu \in (-\pi, \pi]$, $\kappa_1, \kappa_2 > 0$, $\lambda \in (-\infty, \infty)$, and

$$C^{-1} = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_m(\kappa_2). \tag{3.2}$$

The parameters μ and ν determine the mean of the distribution, while κ_1 and κ_2 are precision parameters. The parameter λ determines the nature and strength of association between ϕ and ψ. This density is unimodal when $\lambda^2 < \kappa_1\kappa_2$ and bimodal otherwise. One of the most attractive features of this particular parameterization of the bivariate von Mises is that, when the precision parameters are large and the density is unimodal, it can be well approximated by a bivariate normal distribution with mean (μ, ν) and precision matrix Ω, where $\Omega_{11} = \kappa_1$, $\Omega_{22} = \kappa_2$, and $\Omega_{12} = \Omega_{21} = -\lambda$.

Singh et al. (2002) fit individual sine model distributions to torsion angle datasets. Mardia et al. (2008) developed an extension of the bivariate sine model for n dimensional angular data, but the constant of integration is unknown for $n > 2$, rendering it difficult to use. We instead consider a method based on a Dirichlet process mixture model.

The Dirichlet process, first described by Ferguson (1973), is a distribution of random measures which are discrete with probability one. The Dirichlet process is typically parameterized as having a mass parameter τ_0 and a centering distribution G_0. Using the stick-breaking representation of Sethuraman (1994), a random measure G drawn from a Dirichlet process $DP(\tau_0 G_0)$ takes the form:

$$G(\boldsymbol{B}) = \sum_{j=1}^{\infty} p_j \delta_{\gamma_j}(\boldsymbol{B})$$

where δ_γ is an indicator function equal to 1 if $\gamma \in \boldsymbol{B}$ and 0 otherwise, $\gamma_j \sim G_0$, p'_j $\sim$Beta$(1, \tau_0)$, and $p_j = p'_j \prod_{k=1}^{j-1}(1 - p'_k)$. In this form, the discreteness of G is clearly evident.

This discreteness renders the DP somewhat unattractive for directly modeling continuous data. However it can be effectively used in hierarchical mixture models (Antoniak, 1974). Consider a dataset $z_1, ..., z_n$, and a family of distributions $f(z|\gamma)$ with parameter γ.

A Dirichlet process mixture (DPM) model takes the form:

$$
\begin{aligned}
z_i \mid \gamma_i &\sim f(z_i|\gamma_i) \\
\gamma_i \mid G &\sim G \\
G &\sim DP(\tau_0 G_0)
\end{aligned}
\tag{3.3}
$$

The discreteness of draws from a DP means that there is positive probability that $\gamma_i = \gamma_j$ for some $i \neq j$. For such i and j, z_i and z_j come from the same component distribution, and are viewed as being *clustered* together. The clustering induced by DPM models generates rich classes of distributions by using mixtures of simple component distributions.

While γ is generally taken to be scalar- or vector-valued, there is nothing inherent in the definition of the DP that imposes such a restriction, and more complex centering distributions have been explored (e.g., MacEachern, 2000; De Iorio et al., 2004; Gelfand et al., 2005; Griffin and Steel, 2006; Dunson et al., 2007; Rodríguez et al., 2008). In a model for the distribution of multiple angle pairs, we propose using a hidden Markov model (HMM), a stochastic process, as the centering distribution G_0. In the following section, we describe how to use this hidden Markov model as a component of an informative prior for protein conformation angle data.

3.3 Dirichlet Process Mixture Model for Multiple Alignment Positions

The necessary Bayesian procedures to use a DP mixture of bivariate von Mises sine distributions for modeling torsion angle data at individual sequence positions were developed in Chapter II. In this section we extend this model to multiple sequence positions, and provide a noninformative prior that is directly analogous to the single position model. In addition we describe a method for using an HMM as a centering distribution in an informative prior for sequences of contiguous positions. We also show how to perform density estimation using our model.

Consider a protein family dataset consisting of n angle pair sequences denoted $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Let each observation have m sequence positions, whose angle pairs are denoted $x_{i1}, ..., x_{im}$ for the ith sequence, with $x_{ij} = (\phi_{ij}, \psi_{ij})$. For the moment assume that we have complete data, i.e. that every x_{ij} contains an observed (ϕ, ψ) pair. Then our base model for the jth position in the ith sequence is as follows:

$$\begin{aligned} x_{ij} \mid \theta_{ij} &\sim f(x_{ij} \mid \theta_{ij}) \\ \boldsymbol{\theta}_i \mid G &\sim G \\ G &\sim DP(\tau_0 H_1 H_2), \end{aligned} \tag{3.4}$$

where θ_{ij} consists of the parameters $(\mu_{ij}, \nu_{ij}, \Omega_{ij})$, $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{im})$, and $f(x|\theta)$ is a bivariate von Mises sine model. The distribution G is a draw from a Dirichlet process, while H_1 and H_2 are the centering distributions that provide atoms of the mean and precision parameters, respectively. Note that the product $H_1 H_2$ takes the role of G_0 from (3.3).

For our purposes, H_2 always consists of the product of m identical Wishart distributions we call h_2. This centering distribution assumes independence for the precision parameters of sequence positions given clustering information. Similarly we do not assume a relationship between the precision parameters and the mean parameters for any sequence position, again restricting ourselves to the situation when clustering is known. The use of a Wishart prior for bivariate von Mises precision parameters is motivated by concerns about ease of sampling from the prior distribution and potential issues with identifiability. A more detailed explanation is given in Chapter II.

We discuss two distinct choices for H_1, the centering distribution for the sequence of mean parameters $(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$. The first assumes *a priori* independence of the mean parameters across sequence positions, while the second is designed to share information across adjacent sequence positions using a hidden Markov model based on known properties of protein secondary structure.

3.3.1 Noninformative Prior for Multiple Sequence Positions

A straightforward extension of the existing single position DPM model takes H_1 to be the product of m identical bivariate von Mises distributions we call h_1. For truly noninformative priors, a diffuse von Mises distribution may be replaced by a uniform distribution on $(-\pi, \pi] \times (-\pi, \pi]$. Both the von Mises and uniform versions of the model assume *a priori* independence of the centering parameters (μ_{ij}, ν_{ij}) across sequence positions j. However dependence can still appear in the posterior distribution. While we refer to this as the noninformative model, and use it as such, there is no reason why informative distributions could not be used as the components of H_1, nor must these components be identical. The primary distinguishing feature of this choice of model is that no assumptions are made as to the relationship between the mean parameters at the various sequence positions.

An advantage of this choice for H_1 is that sequence positions j and $j + 1$ need not be physically adjacent in a protein. This situation could be of interest when modeling the joint distribution of amino acid residues which are not neighbors with respect to the primary structure of a protein, but which are close together when the protein is folded.

3.3.2 Informative DPM-HMM Model for Adjacent Sequence Positions

When considering adjacent positions, however, a model assuming independence is not making use of all available information regarding protein structure. For this situation we recommend a centering distribution H_1 that consists of a hidden Markov model incorporating secondary structure information.

We call our model a Dirichlet process mixture on a hidden Markov model space, or DPM-HMM. Hidden Markov models define a versatile class of mixture distributions. An overview of Bayesian methods for hidden Markov models is given by Scott (2002). HMMs are commonly used to determine membership of protein families for template-based structure modeling, but in this case the state space relates to the amino acid sequence, also known as the primary structure (see e.g. Karplus et al., 1997). We propose instead to use an HMM for which the hidden state space consists of the secondary structure type at a particular sequence position. While HMMs incorporating secondary structure have been used for *de novo* structure prediction methods (Boomsma et al., 2008), they have not previously been employed for template-based strategies. We can determine both the transition probabilities between states and the distributions of (ϕ, ψ) angles for each secondary structure type based on datasets in the Protein Data Bank. Such a model provides a knowledge-driven alternative to our noninformative prior from Section 3.3.1 for adjacent sequence positions.

Our model has four hidden states corresponding to four secondary structure metatypes defined by the Definition of Secondary Structure for Proteins (DSSP, Kabsch and Sander, 1983) program: turn (T), helix (H), strand (E), and random coil (C). These four types are condensed from eight basic types, with all helices being characterized as (H), β-turns and G-turns combined into the class (T), and both strands and β-bulges defined as (E). The model for a realization $\boldsymbol{\theta}$ from our hidden Markov model is defined as follows:

$$\begin{aligned} \theta_j \mid s_j \quad &\sim f(\theta_j \mid s_j) \\ s_j \mid s_{j-1} &\sim M(s_j \mid s_{j-1}) \end{aligned}$$

where s_j defines the *state* of the Markov chain at position j, with $s_j \in \{1, 2, 3, 4\}$.
$M(s_j|s_{j-1})$ is a discrete distribution on $\{1, 2, 3, 4\}$ that selects a new state type with probabilities determined by the previous state type. We set our transition probability matrix based on 1.5 million sequence position pairs from the PDB, while the initialization probabilities correspond to the stationary distribution for the chain. Note that $\boldsymbol{s} = (s_1, ..., s_m)$ is an observation from a discrete time Markov process. We then define $f(\theta_j|s_j)$ to be a probability distribution with parameters determined by the current secondary structure state of the chain.

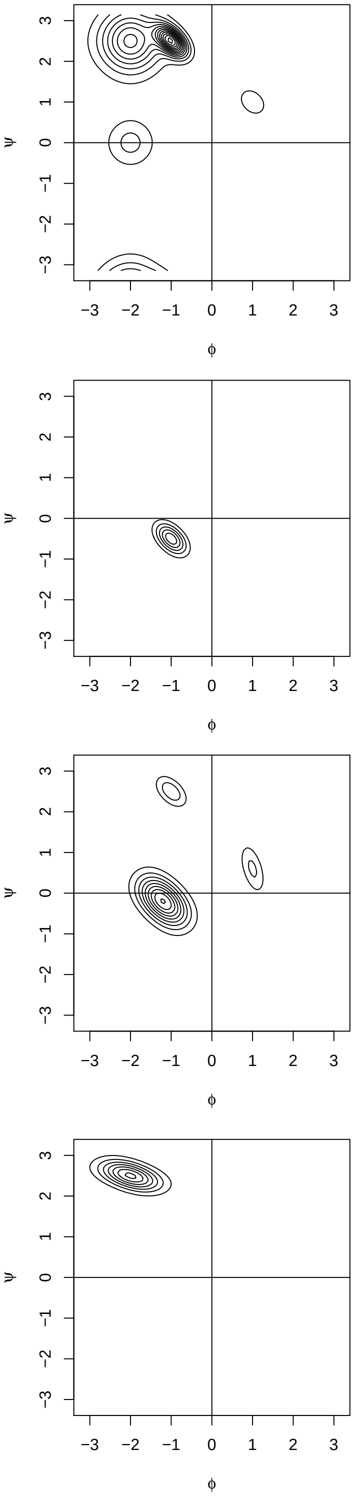
Single bivariate von Mises distributions are not adequate to serve as the state distributions for the four secondary structure types. Instead, we use mixtures of between one and five bivariate von Mises sine models. The amino acids proline and glycine exhibit dramatically different secondary structure Ramachandran distributions, and so were given their own distinct sets of secondary structure distributions. Figure 10 shows the state distributions used for each secondary structure class for the eighteen standard amino acids.

Although these are distributions for the means of the bivariate von Mises distribution, we chose them to mimic the distributions of (ϕ, ψ) angles in each of these secondary structure classes, which means that they are somewhat more diffuse than necessary. The use of these secondary state distributions in conjunction with the Markov chain on the state space allows us to leverage information about secondary structure into improved density estimates, and provides a biologically sound framework for sharing information across sequence positions.

Note that our model is not to be confused with the hidden Markov Dirichlet process (HMDP) proposed by Xing and Sohn (2007). The HMDP is an implementation of a hidden Markov model with an infinite state space, originally proposed by Beal et al. (2002). Their model is an instance of the Hierarchical Dirichlet Process (HDP) of Teh et al. (2006), whereas our DPM-HMM is a standard Dirichlet process with a novel centering distribution.

3.4 Density Estimation

Recall that we are interested in estimating the joint density of $x = (\phi, \psi)$ angles at each sequence position for a candidate structure from some protein family. Our method, as outlined by Escobar and West (1995), involves treating our density estimate as a mixture of components $f(\boldsymbol{x}_{n+1}|\boldsymbol{\theta_{n+1}})$, which in our case are products of bivariate von Mises sine models, mixed with respect to the posterior predictive distribution of the parameters $\boldsymbol{\theta}_{n+1}$.



Coil Prior

p	μ	ν	κ_1	κ_2	λ
0.625	-2.0	2.5	4.00	4.00	0.00
0.208	-1.0	2.5	21.33	21.33	-10.67
0.125	-2.0	0.0	6.25	6.25	0.00
0.043	1.0	1.0	12.21	12.21	-3.66

Helix Prior

p	μ	ν	κ_1	κ_2	λ
1.000	-1.0	-0.5	21.33	21.33	10.67

Turn Prior

p	μ	ν	κ_1	κ_2	λ
0.800	-1.2	-0.2	8.33	8.33	-4.17
0.100	-1.0	2.5	21.33	21.33	-10.67
0.100	1.0	0.6	33.33	8.33	-8.33

Strand Prior

p	μ	ν	κ_1	κ_2	λ
1.000	-2.0	2.5	5.33	21.33	5.33

Figure 10: Graphical and numerical representations of our von Mises mixture distributions for each of the four secondary structure states. Note that this is the general set of secondary structure distributions, and is not used at positions containing the amino acids proline or glycine.

This can be written as:

$$P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \int f(\boldsymbol{x}_{n+1}|\boldsymbol{\theta}_{n+1})dP(\boldsymbol{\theta}_{n+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n). \tag{3.5}$$

This integral cannot be written in closed form, but can be well approximated by Monte Carlo integration. This is achieved by acquiring samples $\boldsymbol{\theta}^1_{n+1}, ..., \boldsymbol{\theta}^B_{n+1}$ from the posterior predictive distribution for $\boldsymbol{\theta}_{n+1}$. Then:

$$\begin{aligned} P(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) &\approx \frac{1}{B}\sum_{k=1}^{B} f(\boldsymbol{x}_{n+1}|\boldsymbol{\theta}^k_{n+1}) \\ &= \frac{1}{B}\sum_{k=1}^{B}\prod_{j=1}^{m} f(x_{n+1,j}|\theta^k_{n+1,j}). \end{aligned} \tag{3.6}$$

While (3.6) can be evaluated for any (ϕ, ψ) sequence $\boldsymbol{x}$, we are typically interested in graphical representations of marginal distributions at each sequence position. For this purpose we evaluate on a 360×360 grid at each alignment position. This general Monte Carlo approach works for joint, marginal, and conditional densities.

3.4.1 Markov Chain Monte Carlo

All that remains is to determine how to obtain the samples from the posterior predictive distribution of $\boldsymbol{\theta}_{n+1}$, which consists of $\boldsymbol{\mu}_{n+1}$, $\boldsymbol{\nu}_{n+1}$, and $\boldsymbol{\Omega}_{n+1}$. Fortunately, while our model is novel, the behaviors of Dirichlet process mixtures, hidden Markov models, and the bivariate von Mises distribution are well understood. The complexity of the posterior distribution prevents direct sampling, but we have developed an effective Markov chain Monte Carlo update scheme using an Auxiliary Gibbs sampler (Neal, 2000).

An initial state can be set by assigning all observations to clusters at random or according to some deterministic method. Examples would be assigning all observations to distinct clusters or assigning all observations to a single cluster. For each initial cluster, draw parameters from the appropriate centering distributions. After the state of our Markov chain has been initialized, our first step is to update the clustering associated with our Dirichlet process. We use the Auxiliary Gibbs sampler of Neal (2000) with one auxiliary component for this purpose. Having updated the clustering, we now must update the parameter values $\boldsymbol{\theta}$ for each cluster by drawing values from full conditional distribution $f(\boldsymbol{\theta}|\boldsymbol{x_c})$, where $\boldsymbol{x_c} = \{\boldsymbol{x}_i : i \in \boldsymbol{c}\}$ and $\boldsymbol{c}$ is the set of indices for members of said cluster. Once again, this distribution is difficult to sample from directly, so we update instead using the full conditional distributions $f(\boldsymbol{\mu}, \boldsymbol{\nu}|\boldsymbol{\Omega}, \boldsymbol{x_c})$ and $f(\boldsymbol{\Omega}|\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{x_c})$.

In the case of the precision parameters $\boldsymbol{\Omega}$, the full conditional density cannot be written in closed form, but is generally well approximated by the Wishart full conditional distribution that results from the assumption that the data have a bivariate normal distribution rather than a bivariate von Mises distribution. We update $\boldsymbol{\Omega}$ by implementing an independence sampler that uses this "equivalent" Wishart distribution as its proposal distribution at each sequence position. Note that under our model, the full conditional distribution of $\boldsymbol{\Omega}$ does not depend on the choice of centering distribution of the mean parameters. The full conditional is proportional to:

$$\begin{aligned} L(\boldsymbol{\Omega}|\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{x_c}) &\propto H_2(\boldsymbol{\Omega})\ L(\boldsymbol{x_c}|\boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\ &= \prod_{j=1}^{m} h_2(\Omega_j) \prod_{i \in \boldsymbol{c}} f(x_{ij}|\mu_j, \nu_j, \Omega_j) \end{aligned} \tag{3.7}$$

where h_2 is our component Wishart prior for a single sequence position, and f is a bivariate von Mises sine model with the relevant parameters. Notice that the positions are independent given the clustering information, so it is trivial to update each Ω_j separately.

After updating the precision parameters at each sequence position, we proceed to update $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ using an independence sampler. For our noninformative prior, with a centering distribution consisting of a single sine model, we use the update method described in Chapter II. In this case, with $H_1 = (h_1)^n$ where h_1 is a bivariate von Mises distribution, the full conditional distribution is proportional to:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\nu} | \boldsymbol{\Omega}, \boldsymbol{x_c}) &\propto H_1(\boldsymbol{\mu}, \boldsymbol{\nu}) \; L(\boldsymbol{x_c} | \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\ &= \prod_{j=1}^{m} h_1(\mu_j, \nu_j) \prod_{i \in \boldsymbol{c}} f(x_{ij} | \mu_j, \nu_j, \Omega_j) \end{aligned} \tag{3.8}$$

The DPM-HMM case where H_1 is defined to be a hidden Markov model is somewhat more complicated. The positions are no longer *a priori*, and therefore *a posteriori*, independent given the clustering information. However, if the state chain $\boldsymbol{s}$ is known, draws from the full conditional are trivial. Therefore we rewrite our full conditional distribution, which is proportional to:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\nu} | \boldsymbol{\Omega}, \boldsymbol{x_c}, \boldsymbol{s}) &\propto H_1(\boldsymbol{\mu}, \boldsymbol{\nu} | \boldsymbol{s}) \; L(\boldsymbol{x_c} | \boldsymbol{\Omega}, \boldsymbol{\mu}, \boldsymbol{\nu}) \\ &\propto \prod_{j=1}^{m} f(\mu_j, \nu_j | s_j) \prod_{i \in \boldsymbol{c}} f(x_{ij} | \mu_j, \nu_j, \Omega_j) \end{aligned} \tag{3.9}$$

where $f(\mu, \nu | s_j)$ is the secondary structure based distribution determined by the state at position j. Recall that our priors are finite mixtures of bivariate von Mises sine distributions. Thus if we can generate draws from the full conditional distribution of $\boldsymbol{s}$, we can update μ_i and ν_i at each sequence position much as we did before. We use the forward-backward (FB) algorithm of Chib (1996) to sample the full conditional distribution of $\boldsymbol{s}$. Note that $\boldsymbol{s}$ given $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is independent of the data. Once we have the state information, generating samples from the distributions $\mu_j, \nu_j | s_j, \Omega_j, x_{\boldsymbol{c}j}$ is a straightforward process using an independence sampler, the details for which are given in Appendix B.

An outline of the complete MCMC procedure is given in Figure 11.

3.4.2 The Sparse Data Problem

The model as described up to this point does not fully account for the complexity of actual protein alignment data. Rather than being a simple vector $\boldsymbol{x}_i$ of bivariate (ϕ, ψ) observations, the real data also includes a vector $\boldsymbol{a}_i$ of length m which consists of variables

1. Initialize the parameter values:
   (a) Choose an initial clustering. Two obvious choices are: (1) one cluster for all of the angle pair sequences, or (2) each angle pair sequence in a cluster by itself.
   (b) For each initial cluster $\boldsymbol{c}$ of observed angle pair sequences, initialize the value of the common bivariate von Mises parameters $\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\Omega}$ by sampling from the centering distribution $H_1(\boldsymbol{\mu}, \boldsymbol{\nu})H_2(\boldsymbol{\Omega})$ of the DP prior.
      i. For the noninformative prior model, sample from each of m independent von Mises and Wishart distributions.
      ii. For the DPM-HMM, obtain initial values for $\boldsymbol{\Omega}$ from m independent Wishart distributions and $\boldsymbol{\mu}, \boldsymbol{\nu}$ from the hidden Markov model.
2. Obtain draws from the posterior distribution by repeating the following:
   (a) Given the mean and precision values, update the clustering configuration using one scan of the Auxiliary Gibbs sampler of Neal (2000).
   (b) Given the clustering configuration and mean values, update the precision matrix Ω for each sequence position in each cluster using the Wishart independence sampler described in Chapter II.
   (c) If using the DPM-HMM, obtain a draw from the full conditional distribution of the state sequence $\boldsymbol{s}$ using the FB algorithm developed by Chib (1996) for each cluster.
   (d) Given the clustering configuration, precision values, and (if applicable) state information, update the values of (μ, ν) for each sequence position in each cluster using the independence sampler given in Appendix B.

Figure 11: Computational procedure for DPM-HMM and nonparametric prior models for torsion angle pairs.

indicating whether or not peptide i was observed at each sequence position. Let $a_{ij} = 1$ if peptide i is included at alignment position j, and 0 otherwise. This data structure is unique in several ways. Notice that $\boldsymbol{a}_i$ is not only known for proteins with solved structure, but is also typically available for a target peptide sequence. Therefore, we can avoid fitting a model that includes alignment positions which are not of interest for our particular problem. Secondly, this is not a true "missing data" problem, as the unobserved sequence positions are not only absent from our dataset, but do not exist.

Our model is able to adjust to sparse data with the following modification. Recall that the full conditional distributions could be divided up into a prior component and a data component at each sequence position. This makes it trivial to exclude an observation from the likelihood, and hence posterior distribution calculation, at sequence positions where it is not observed. For example, we can modify the full conditional distribution of the means in the DPM-HMM model, given in equation (3.9), to be:

$$f(\boldsymbol{\mu}, \boldsymbol{\nu} | \boldsymbol{\Omega}, \boldsymbol{x_c}, \boldsymbol{s}) \propto \prod_{j=1}^{m} f(\mu_j, \nu_j | s_j) \prod_{i \in \boldsymbol{c}} f(x_{ij} | \mu_j, \nu_j, \Omega_j)^{a_{ij}}$$

The full conditional distributions for the precision parameters and the means with a noninformative prior, equations (3.7) and (3.8) respectively, can be modified in a similar manner. The likelihood of $\boldsymbol{x}_i | \boldsymbol{\theta}$ is also used by the Auxiliary Gibbs sampler. Once again, adjust to absent data by removing unobserved positions from the likelihood.

This model provides a straightforward method to cope with the sparse data problem inherent in protein structure prediction. Note that the situation in which there is ample data generally but sparse data at a few sequence positions particularly highlights the value of the DPM-HMM model. Secondary structure at a sparse position can be inferred based on the surrounding positions, which can allow us to provide a better density estimate at positions with few observed data points.

3.5 Application to Loop Modeling in the Globin Family

3.5.1 Background

A protein's fold, or tertiary structure, consists of multiple elements of local, regular secondary structure (repeating local motifs) connected by the more variable loops and turns of various lengths. These loop and turn regions can be vital to understanding the function of the protein, as is the case in the immunoglobulin protein family where the conformation of the highly variable loops determine how an antibody binds to its target antigens to initiate the body's immune response. These loop regions also tend to be the most structurally variable parts of the protein, and modeling their structure remains an outstanding problem in protein structure prediction (Baker and Sali, 2001). Current knowledge-based loop modeling methods draw on generic loop libraries. Library-based methods search the Protein Data Bank for loops with entrance and exit geometries similar to those of the target loop, and use these PDB loops as templates for the target structure (e.g. Michalsky et al., 2003). Note that library-based methods differ from typical template-based modeling in that they do not confine themselves to loops within the target protein's family. Strictly within family estimates have not previously been possible. Using the DPM-HMM model, we are able to compare a library-based approach to a purely within family template-based method for the EF loop in the globin family.

The globins are proteins involved in oxygen binding and transport. The family is well studied and has many known members. Therefore, the globin fold is suitable as a test case for template-based structure prediction methods. A globin consists of eight helices packed around the central oxygen binding site and connected by loops of varying lengths. The helices are labeled A through H, with the loops labeled according to which helices they connect. The EF loop is the longest loop in the canonical globin structure. We generated a simultaneous alignment of 94 members of the globin family with known tertiary structure

Table 1: A table giving the details on the EF loop for an alignment of 94 members of the globin family. The columns are the alignment position, the number of proteins represented at the position, the most conserved amino acid(s) at the alignment position, and the total number of distinct amino acids observed at the alignment position.

Position	# of Proteins	Most Conserved AA	# of AAs
93	94	LEU	7
94	94	ASP	10
95	94	ASN	9
96	26	ALA	11
97	28	GLY	8
98	28	LYS	10
99	94	LEU	7
100	1	THR	1
101	2	VAL	1
102	2	THR ARG	2
103	93	LYS	13
104	94	GLY	15
105	94	ALA	15
106	94	LEU	10

using MUSCLE (Edgar, 2004). For this alignment, positions 93-106 correspond to the EF loop.

Table 1 gives a summary of the behavior of 94 representative globins in the EF loop region. There is considerable diversity in both the length and amino acid composition of this loop. Representative loops were between 8 and 14 amino acids long, and the highly conserved regions, particularly at the tail end of the loop, exhibited considerable variability in amino acid composition.

We compare three different methods for loop modeling: our DPM-HMM method with globin family data, the noninformative prior model with globin family data, and a library-based approach. Library approaches generate lists of loops similar to the target and use these as templates for the target loop, generating a discrete distribution which almost surely has mass 0 at the true conformation of the unknown loop. To make this method comparable to our density-based approaches, we used our noninformative prior model on library datasets to generate a continuous density estimate. Note that all sequences in a library dataset are of the same length, which means that they will never exhibit sparsity. For this reason, fitting the DPM-HMM model on the library dataset would not present much improvement over the noninformative model.

3.5.2 Parameter Settings

For each of the 94 globins in the alignment we generated density estimates using each of the three methods in question. For the DPM-HMM and noninformative models, we excluded the target from the dataset used to generate the density estimates, but used amino acid and sparse data information from the target protein. This is reasonable since primary structure based alignments are available for template modeling of an unknown protein. For the library-based estimate, we applied our noninformative prior model sequences from the coil library of Fitzkee et al. (2005) which have the same length as the target sequence, and have at least four sequence positions with identical amino acids. Library datasets ranged in size from 17 to 436 angle pair sequences.

For each of our models we ran two chains: one starting with all observations in a single cluster and one with all observations starting in individual clusters. Each chain was run for 11,000 iterations with the first 1,000 being discarded as burnin. Using 1 in 20 thinning, this gave us a combined 1,000 draws from the posterior distribution of the parameters.

In all cases, our Wishart prior used $v = 1$, and we set the scale matrix B to have diagonal elements of 0.25 and off-diagonal elements of 0. Note that we use the Bernardo and Smith (1994, pp. 138–139) parameterization, with an expected value of $vB^{-1} = B^{-1}$. Our choice of v was motivated by the fact that this is the smallest possible value for which moments exist for the Wishart distribution, and higher values would have lead to a more informative prior. The choice of B gave an expected standard deviation of about 30 degrees and assumed *a priori* that there was no correlation between ϕ and ψ, which seemed to work well in practice. For our noninformative prior on the means, we took h_1 to have $\mu_0 = \nu_0 = 0$, $\kappa_{10} = \kappa_{20} = 0.1$, and $\lambda_0 = 0$. This provided a diffuse centering distribution.

In all cases we took the DP mass parameter τ_0 to be 1, although our results were robust to departures from this value. For example, for two randomly selected proteins we gave values for τ_0 ranging between 0.2 and 15 giving prior expected numbers of clusters from approximately 2 to 30. For our first peptide the observed mean cluster number ranged from 3.96 to 4.46, while the second had values from 4.40 to 4.65. Thus even our most extreme choices for the mass parameter changed the posterior mean number of clusters by less than 1.

3.5.3 Results of Comparison to Library

We performed pairwise comparisons for each of our models using the Bayes factor, defined as:

$$B((\boldsymbol{\phi}, \boldsymbol{\psi})) = \frac{f((\boldsymbol{\phi}, \boldsymbol{\psi})|M_1)}{f((\boldsymbol{\phi}, \boldsymbol{\psi})|M_2)}$$

where M_1 and M_2 are density estimates generated by two of our three possible models. We present the results of the analyses for our 94 leave-one-out models in Table 2.

First we will address the comparison between the DPM-HMM and noninformative models using the globin data. These models show far more similarity to each other than to the noninformative model using the library data, both in terms of the number of Bayes factors indicating superiority on each side, and the fact that those Bayes factors tended to be smaller in magnitude than those generated by comparisons to the library models. Indeed, at positions with more than 30 observations the marginal distributions generated by the two models appear to be very similar. In total, the DPM-HMM model was superior to the noninformative prior model in 59 out of 94 cases. For the null hypothesis that the probability that the DPM-HMM is superior to the noninformative model is less than or

Table 2: Comparison between the DPM-HMM model on the globin family data, noninformative prior with globin data, and noninformative model with library data. The columns Model X to Model Y give the percentage of the time that the likelihood for the target conformation using Model X was greater than the likelihood of the same conformation using Model Y. This is the equivalent to a Bayes factor comparison with Model X in the numerator being greater than 1.

Loop Length	Total	DPM-HMM to Library	Noninf to Library	DPM-HMM to Noninf
8	66	100%	100%	70%
10	3	67%	67%	67%
11	23	100%	96%	39%
13	1	100%	100%	100%
14	1	100%	100%	100%
All	94	99%	98%	63%

equal to 0.5, a binomial test yields a p-value of 0.009. Of these Bayes factor results, 68 met standard criteria for substantial evidence of superiority ($|\log_{10}(B)| > 1/2$, Kass and Raftery, 1995), of which 45 supported the use of the DPM-HMM model, giving a p-value of 0.005. This is in addition to the fact that the combined Bayes factor, the product of all of the individual comparisons, has a value of 10^{38}, which provides overwhelming evidence in favor of using the DPM-HMM rather than the noninformative model. For this reason we will only refer to the DPM-HMM model when making use of the globin dataset for the remainder of the chapter.

Recall that the library model made use of loops which were of the same length as the target and had a certain degree of similarity in terms of amino acid sequence. Thus the coil library does not exhibit any sparse data behavior. It is also unlikely to recapture the globin family EF loops due to the considerable variability in both length and amino acid composition. Our results indicate that the DPM-HMM model overwhelmingly outperforms the library-based method. Not only is the relevant Bayes factor greater than 1 in 93 out of 94 cases, it is greater than 100 in 92 cases. The case in which the library-based method outperformed the DPM-HMM was also significant according to the Kass and Raftery (1995) criteria, so there were no ambiguous individual cases. The combined Bayes factor was 10^{959}, indicating that the DPM-HMM model was definitely superior to the library overall.

Figure 12 shows marginal density estimates generated for prototypical globin "1t1nB" for both models, along with the true (ϕ, ψ) sequence for the protein for a portion of the EF loop. By searching the PDB for loops that are similar to the target in terms of length and sequence identity, the library method tends to place considerable mass in areas of conformational space that are not occupied by members of the globin family. While the members of the dataset for the globin family may not match the target loop in terms of length or amino acid sequence, by virtue of being globins themselves they provide a better match to the target conformation. This pattern of improvement held true regardless of loop length. Significant improvement was found even for the length 13 and 14 loops, for which sparse data was a particular problem.

Figure 13 shows the effect of the hidden Markov model prior as the number of observations increases. The density estimates shown are the DPM-HMM and noninformative prior model fits for the globin "1d8uA," for which the EF loop is of length 14. This means that for the leave-one-out fit there is a sequence position with no observed data. At this position, the DPM-HMM model clearly shows the influence of the coil state distribution, while the noninformative model gives a distribution which is close to uniform. As the number of observations at a given sequence position increases, the differences between the two models become less significant. For 25 data points the densities are very similar, and for 92 they are practically identical. This behavior is desirable, as we wish for the DPM-HMM to compensate when data is very sparse but not to overwhelm a large number of observations.

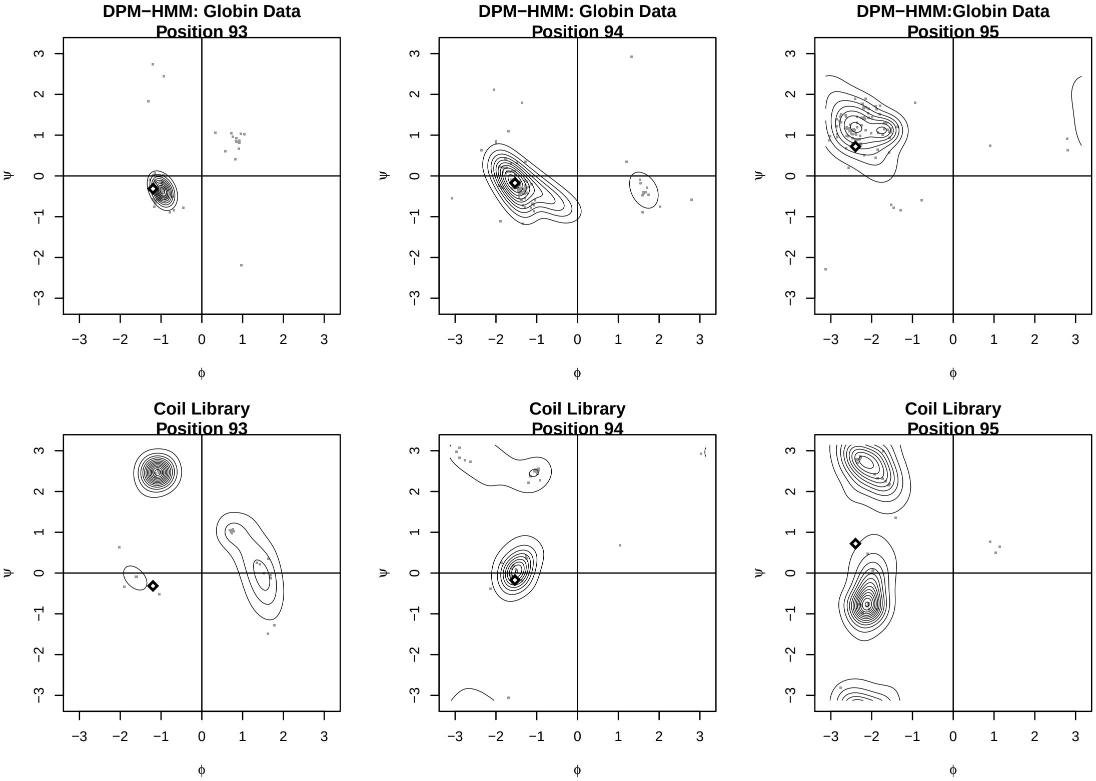


Figure 12: Density estimates for positions 93, 94, and 95 for protein “1t1nB.” The gray dots indicate the data used to fit the model, while the diamonds show the true (ϕ, ψ) conformation of the target protein.

3.5.4 *Results of Comparison to DBN-torus*

In addition to comparing the DPM-HMM to the knowledge-based library method, we have also conducted a comparison to the *de novo* DBN-torus sequence prediction method of Boomsma et al. (2008). Unlike the previously addressed library-based methods, DBN-torus uses continuous density estimates, but is not customized for loop regions. It can be used to generate sequences of predicted angle pairs given amino acid data, secondary structure data, or no input at all. The best results for DBN-torus are generated using amino acid data and predicted secondary structure data. For each of our 94 targets, we generated 1,000 candidate draws using the DPM-HMM, DBN-torus with predicted secondary struc-

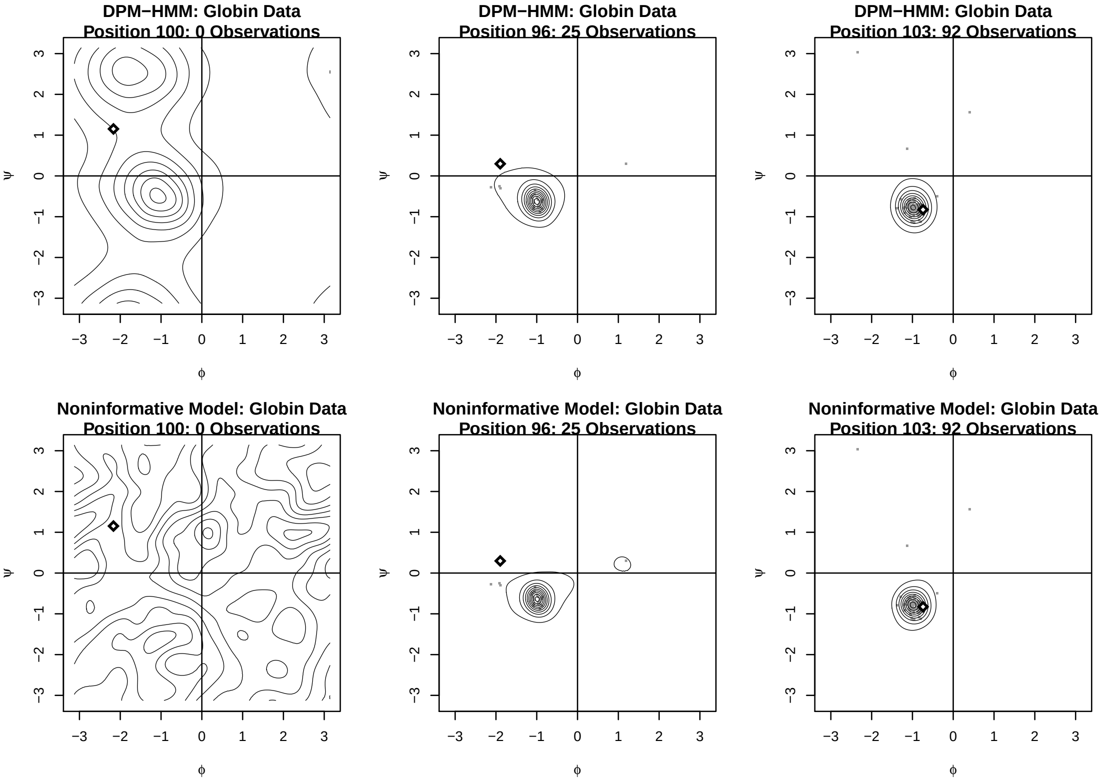


Figure 13: A comparison of the DPM-HMM and noninformative prior models for the length 14 loop of the globin “1d8uA.” Again, the gray dots represent data used to generate the density estimate, while the diamonds indicate the true (ϕ, ψ) values.

ture data from PsiPred (McGuffin et al., 2000), and DBN-torus using the true secondary structure data. Although having exact knowledge of secondary structure for a target protein is unrealistic in practice, it gives an idea of how well DBN-torus can perform with optimal secondary structure prediction. We followed the strategy of Boomsma et al. (2008) of using the angular RMSD to judge the accuracy of our predictions. For our purposes, the angular RMSD is defined as:

$$aRMSD((\boldsymbol{\phi}_1, \boldsymbol{\psi}_1), (\boldsymbol{\phi}_2, \boldsymbol{\psi}_2)) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\Delta\phi_i^2 + \Delta\psi_i^2)}$$

where $\Delta z_i = \min(|z_{1i} - z_{2i}|, 2\pi - |z_{1i} - z_{2i}|)$.

For each target, the best candidate judged by minimum aRMSD was selected, and the

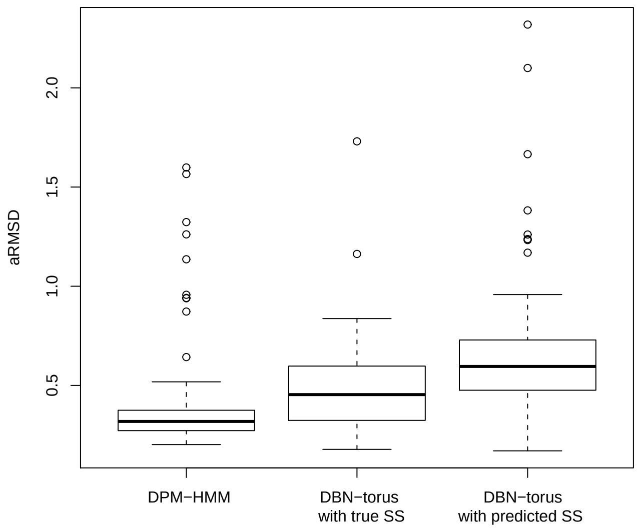


Figure 14: Comparison of prediction accuracy between the DPM-HMM and DBN-torus. DBN-torus has been given either predicted or real secondary structure information as input. Small aRMSD values, here given in radians, indicate predictions which are close the target's true tertiary structure.

results are summarized in Figure 14. The DPM-HMM provides a better minimum aRMSD estimate than DBN-torus in 75/94 cases with predicted secondary structure information and 67/94 cases with true secondary structure information. Note that even under this best case scenario the DPM-HMM tends to provide better predictions than does DBN-torus. This is unsurprising, as template-based methods typically outperform *de novo* methods where a template is available. Proteins for which DBN-torus outperforms our DPM-HMM method often contain an EF loop whose conformation is not a close match to other members of the globin family. In such cases, good conformations are more likely to be sampled from DBN-torus, which is based on the entire PDB, rather than the DPM-HMM mimicking the behavior of the other globins.

3.6 Discussion

We have presented a novel model for protein torsion angle data that is capable of estimating the joint distribution of up to around 15 angle pairs simultaneously, and applied it to extend template-based modeling to the notoriously difficult loop and turn regions. In contrast to existing methods such as library-based loop prediction and DBN-torus, our model is designed to make use of only data from highly similar proteins, which gives us an advantage when such data is available. This is a significant advance in terms of statistical models for this type of data, as well as a new approach to template-based structure prediction. In addition to providing the basic model, we proposed two possible prior formulations with interesting properties.

Our noninformative prior model, which is the direct extension of the single position model from Chapter II, provides a method to jointly model sequence positions which may or may not be adjacent in terms of a protein's primary structure. This model allows for the estimation of joint and conditional distributions for multiple sequence positions, which permits the use of innovative methods to generate candidate distributions for protein structure.

While the noninformative prior model represents a significant advance over existing methods, we also present an alternative model that incorporates prior information about protein structure. This DPM-HMM model, which uses a hidden Markov model as the centering distribution for a Dirichlet process, uses the unique characteristics of a protein's secondary structure to generate superior density estimates for torsion angles at sequential alignment positions. We use a Bayes factor analysis to demonstrate that density estimates generated with this model are closer to the true distribution of torsion angles in proteins than our alternative ignoring secondary structure.

Regardless of our prior formulation, the model is capable of accommodating the sparse data problem inherent in protein structural data, and in the case of the DPM-HMM can leverage information at adjacent sequence positions to compensate for sparse data. This allows, for the first time, the extension of template-based modeling to the loop regions in proteins. We show that within family data provides superior results to conventional library and PDB-based loop modeling methods. As loop modeling is one of the critical problems in protein structure prediction, this new model and its ability to enhance knowledge-based structure prediction represents a significant contribution to this field.

Recall that our model treats the parameters of the bivariate von Mises sine model nonparametrically through the use of the Dirichlet process prior centered on a parametric distribution. It is a matter of some interest to compare this to the parametric alternative of using the centering distribution itself as the prior for the bivariate von Mises parameters. This would be equivalent to limiting our model to a single mixture component. Although not every sequence position gives a strong indication of multiple mixture components, there is at least one such sequence position for every loop in our dataset. (See, for example, position 94 for the coil library dataset in Figure 12.) Attempts to model this data using only a single component distribution lead to poor results, particularly since our model enforces unimodality for each component via the Wishart prior. While the HMM prior does allow for a mixture of bivariate von Mises distributions, all of these components will converge to the same distribution as the number of observations increases, effectively reducing us to a single component model again. The inadequacy of such a single component model is reflected in the strong preference of the data for multiple clusters. While the prior expected number of clusters goes to 1 as the mass parameter τ_0 goes to 0, we found that the posterior mean number of clusters only decreased by 1 (typically from 4 to 3) when τ_0 decreased from 1 to 10^{-10}.

In working with our sampling schemes for both the DPM-HMM and noninformative prior models we did occasionally encounter slow mixing and convergence problems, particularly as the number of sequence positions under study increased. Figure 15 shows the effects on the total number of clusters and entropy (Green and Richardson, 2001) per iteration caused by increasing sequence length. As the number of positions under study increases, there is a greater chance of getting stuck in particular conformations, and also a subtler tendency towards having fewer observed clusters. Although in this example the effects are fairly mild, more severe issues can occur even at relatively short sequence lengths. However, even when problems appear to be evident on plots of standard convergence diagnostics, the density estimates generated by separate chains can be quite similar. For this reason we recommend comparing the density estimates generated by multiple chains in addition to the standard methods of diagnosing convergence problems.

We do not recommend that our method be used for simultaneous modeling of more than about 15 sequence positions and convergence diagnostics should always be employed. The use of multiple MCMC chains with different starting configurations is also highly encouraged. Particular care should be taken with the noninformative prior model, which seems to be more prone to these sorts of problems. We did not observe any effect of sparse data on the speed of convergence or mixing.

Increases in sequence length and sample size both increase run time for our software, although sequence length is the primary practical restriction as protein families tend to have fewer than 100 members. For the analysis of the full globin dataset with 5, 10, 15, or 20 sequence positions, the run times for two chains with 11,000 iterations using a 3 GHz processor were between 1 hour and 3.5 hours for the noninformative model and 2 hours to 8 hours for the DPM-HMM.

As the emphasis in this chapter is on loop modeling, which by its very nature is limited to contiguous sequence positions, our application does not reflect the full extent of the flexibility of our model. Our general method is a good source of simultaneous continuous density estimates for large numbers of torsion angle pairs. This allows us to generate

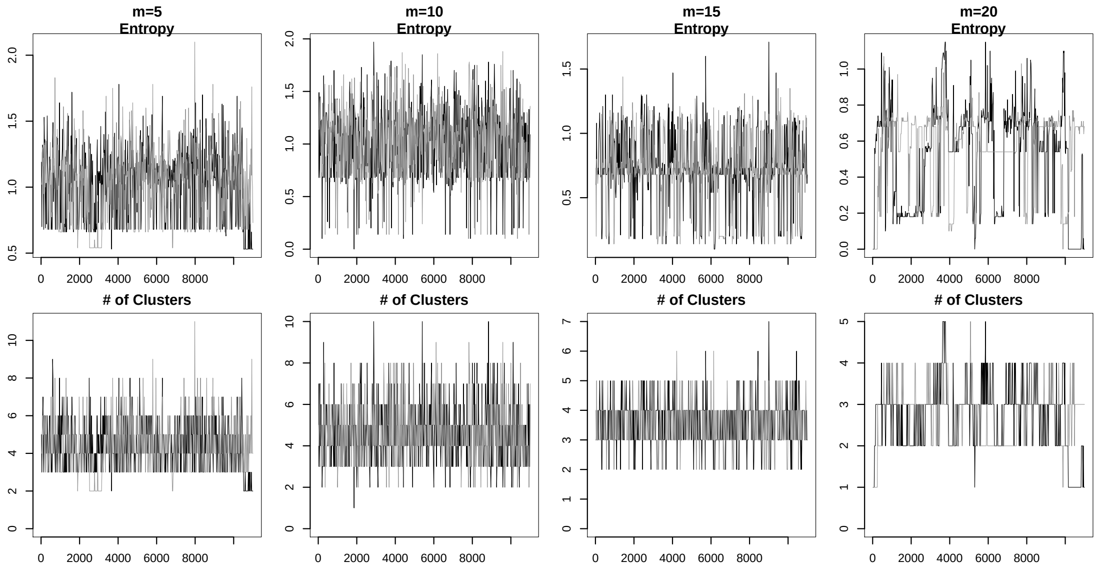


Figure 15: Convergence diagnostics for density estimates using the noninformative prior model on the globin data with contiguous sequences beginning at position 93. Notice how mixing worsens as the number of sequence positions increases.

candidate models by sampling from joint distributions, or to propagate a perturbation of the torsion angle sequence at a single position up and down the chain through the use of conditional distributions. Our noninformative prior model, while less impressive than the DPM-HMM for contiguous sequence positions, can be applied to far richer classes of torsion angle sets. This allows the modeling of the behavior of tertiary structure motifs, which are composed of amino acids which are not adjacent in terms of primary structure, but which are in close contact in the natural folded state of a protein. It can even be used to investigate the structure of polypeptide complexes, as the (ϕ, ψ) positions modeled are not required to belong to the same amino acid chain. The ability to model large numbers of (ϕ, ψ) pairs simultaneously is an exciting advance which will offer new avenues of exploration for template-based modeling, even beyond the field of loop prediction.

The software used in this analysis is available for download at http://www.stat.tamu.edu/∼dahl/software/cortorgles/.

CHAPTER IV

A BAYESIAN NONPARAMETRIC MODEL FOR MULTIVARIATE DEPENDENCE

4.1 Introduction

In the two previous chapters we presented techniques for modeling the protein backbone in the form of (ϕ, ψ) torsion angle pairs. However, the protein backbone is not in itself a complete picture of the structure of a protein. In addition to the four heavy backbone atoms, each amino acid from a protein sequence has a unique side-chain structure. These side-chains are different for all 20 naturally-occurring amino acids, and these differences ultimately determine the final three-dimensional structure of a protein. Our probabilistic description of protein structure is not complete without including information on side-chain location.

While side-chains can consist of many atoms, we will represent a side-chain's position with the location of the side-chain center of mass in three-dimensional space. Under this model, we have five variables of interest at each sequence positions: the (ϕ, ψ) backbone angle pair and the (x, y, z) coordinates of the centroid. This model is more complex than it first appears, as it calls for a joint distribution between two angular and three linear random variables. We saw in Chapters II and III that simple elliptical angular distributions are often insufficient for modeling torsion angle pairs, so some kind of mixture model is appropriate. Unfortunately, existing joint models for combinations of angular and linear variables are not amenable to mixture modeling due to the lack of a closed form solution for the constant of integration in higher dimensions (Johnson and Wehrly, 1978). However, torsion angles and centroids could easily be described marginally in terms of mixtures of bivariate von Mises and trivariate normal distributions respectively. We are left with two relatively simple marginal models, but no clear way to link them in a joint distribution.

A standard approach to these kinds of problems has been the use of copulas. Such methods allow for marginal distributions to be specified first, and then their dependence can be modeled via a function called a copula. Nelsen (2006) provides a good introduction to copulas and their properties. Parametric copula models, in which both the marginal distributions and copula function are members of parametric families, are generally used in practice. However, we are interested in the situation when the marginal distributions of interest and their dependence relationship may not necessarily be well described by a parametric family. There has been work with semiparametric copula models, for which the marginal distributions are estimated nonparametrically but the copula is chosen from a parametric family (Genest et al., 1995), and fully nonparametric models for which both copula and marginals are nonparametrically estimated (Chen and Huang, 2005). However, these nonparametric copulas methods have not been extended to multivariate marginals. Even the use of standard parametric copulas is complicated by potential compatibility problems for multivariate marginal distributions (Nelsen, 2006, pp. 105–108).

We propose an alternative to nonparametric copula models which we call a Dirichlet process dependence model (DPDM). Our motivation is to develop a general Bayesian framework for multivariate problems which are relatively simple for either univariate or multivariate marginal modeling, but which present difficulties in the context of the complete multivariate model. Such situations could arise when a joint multivariate distribution does not exist for some data structure, or when such distributions are not well suited for mixture modeling. Note that we are considering cases when we are not necessarily interested in quantifying dependence, and in fact simple numeric measures of dependence may not exist. Our proposed model incorporates dependence information purely through clustering induced by a Dirichlet process. Models fitting this description were proposed previously in Chapter III for angle pair sequences and by Dunson and Xing (2009) for sets of categorical random variables. However, we present a more general framework for such models which allows the experimenter to select their marginal distribution types to suit their particular problem. In contrast with previous work, we do not require the marginal variables to share a distribution type. The DPDM may be fully nonparametric or include marginal information in the form of a prior. It can accommodate any marginal distribution (continuous, discrete, or categorical) which can be modeled with standard DP mixture modeling techniques, and, in addition to permitting the joint modeling of disparate data types, the DPDM can also accommodate component marginals of any dimension.

In Section 4.2, we present both the model formulation for the DPDM and a general outline for computation. Using strategies developed for Bayesian nonparametric density estimation, we are able to develop joint density functions which allow for straightforward joint, marginal, and conditional computation and sampling. Section 4.3 explores some of the properties of the DPDM model using two-dimensional examples. In Section 4.4 we use the DPDM to develop joint density estimates for torsion angle pairs and centroids for a protein structure dataset. In Section 4.5 we discuss our conclusions about both the protein data analysis and the properties of our model.

4.2 The Dirichlet Process Dependence Model

4.2.1 Basic Model Formulation

The Dirichlet process (DP) is a distribution on almost surely discrete distributions first described by Ferguson (1973). While the discreteness of draws from a Dirichlet process makes them unsuitable as models for continuous data, the Dirichlet process works well when incorporated into a prior for mixture modeling (Antoniak, 1974). Dirichlet process mixture (DPM) models for density estimation are described by Ferguson (1983) for the univariate case and Tiwari et al. (1988) for a general multivariate density. Our model takes the basic DPM framework, and adjusts it to allow for the specification of specific marginal distribution types. We then take advantage of DP clustering to develop joint density estimates incorporating the specified marginal structure.

Consider a set of vector valued observations $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$ where each observation $\boldsymbol{x_i}$ has components $x_{i1}, ..., x_{im}$. Say that we have a partition $\{p_1, ..., p_r\} = \boldsymbol{P}$ on the integers 1 to m such that for each set $\boldsymbol{x}_{ip_j} = \{x_{il}\}_{l \in p_j}$, we have a suitable Bayesian marginal model. By this we mean that we can model $\boldsymbol{x}_{1p_j}, ..., \boldsymbol{x}_{np_j}$ as coming from a mixture of parametric distributions $g_j(\cdot|\theta_j)$ with suitable parameter θ_j. In order to satisfy the computational requirements of our model, we will require a prior distribution $H_j(\theta_j)$ from which we can obtain samples directly. Note that we do not require that the distributions g_j belong to the same parametric family, or even be of the same dimension. Nor do we require that the parameter θ_j be univariate. We define a Dirichlet process dependence model for this data as follows:

$$\begin{aligned} \boldsymbol{x}_{ip_j} |\, \theta_{ij} &\sim g_j(\boldsymbol{x}_{ip_j} | \theta_{ij}) \\ \boldsymbol{\theta_i} |\, G &\sim G \\ G &\sim DP(\tau_0 \prod_{j=1}^{r} H_j) \end{aligned}$$

where G is a draw from a Dirichlet process, τ_0 is the DP mass parameter, and the centering distribution treats all of the component parameters within a draw as being independent. Under this model, a set of parameters $\boldsymbol{\theta_i}$, which defines all marginal components g_j for a given observation $\boldsymbol{x}_i$, is distributed according to G, a draw from a Dirichlet process. Since G is almost surely discrete, there is a positive probability that $\boldsymbol{\theta}_i = \boldsymbol{\theta}_l$ for some $l \neq i$. If this is the case, then we consider $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$ to be clustered together. The centering distribution component H_j for each parameter θ_{ij} is taken to be the prior distribution from the marginal case. Given clustering information we consider θ_{ij} and θ_{ik}, and thus x_{ip_j} and x_{ip_k}, to be independent for $j \neq k$.

Note that the noninformative prior model from Chapter III took this form, where our component distributions were bivariate von Mises and H_i was the product of a bivariate von Mises distribution and a Wishart distribution. This type of model was also employed by Dunson and Xing (2009) for multivariate unordered categorical data. While in both of these cases the marginal distributions all share a parametric family, this is not necessary.

4.2.2 Computation

We will use the method described by Escobar and West (1995) to generate a density estimate for $\boldsymbol{x}_{n+1}$. Specifically, we must obtain B draws from the posterior distribution $\boldsymbol{\theta}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Given those draws, our density estimate for $\boldsymbol{x}_{n+1}$ is:

$$f(\boldsymbol{x}_{n+1}|\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \frac{1}{B}\sum_{i=1}^{B}\prod_{j=1}^{r} g_j(\boldsymbol{x}_{n+1,p_j}|\theta_j^{(i)})$$

where $\theta_j^{(i)}$ is the jth component of the ith draw from the posterior distribution of $\boldsymbol{\theta}$. Notice that this is a finite mixture distribution with the property that our components $\boldsymbol{x}_{ip_j}$ and $\boldsymbol{x}_{ip_l}$, $l \neq j$ are independent given the knowledge of which mixture component they are drawn from. This allows us to derive the DPDM marginal density estimate for component $\boldsymbol{x}_{n+1,p_j}$:

$$f(\boldsymbol{x}_{n+1,p_j}) = \frac{1}{B}\sum_{i=1}^{B} g_j(\boldsymbol{x}_{n+1,p_j}|\theta_j^{(i)})$$

and the conditional distribution:

$$f(\boldsymbol{x}_{n+1,p_j}|\{\boldsymbol{x}_{n+1,p_l}\}, l \in \mathcal{L}) = \sum_{i=1}^{B} w_i g_j(\boldsymbol{x}_{n+1,p_j}|\theta_j^{(i)})$$

where $w_i \propto \prod_{l\in\mathcal{L}} g_l(\boldsymbol{x}_{n+1,l}|\theta_l^{(i)})$, and $\sum_{i=1}^{B} w_i = 1$. Notice that our estimated marginal and conditional distributions are finite mixtures of our component modeling distribution g_j. Of particular interest is the fact that the conditional distribution for any component is simply a reweighted version of the marginal distribution. This gives us straightforward methods for conditional computation and sampling.

However, all of these density estimates require that we first sample from the full conditional distribution of $\boldsymbol{\theta}$. This distribution will generally not be tractable, so we provide a general algorithm involving full conditional distributions which should work for any choice of components g_j for which marginal Bayesian modeling is possible. Posterior computation should proceed in an iterative fashion: first update the clustering of the observations given cluster parameter values, and then update all parameter values in a parametric fashion within the clusters.

A good review of options for full conditional updates of data clustering can be found in Neal (2000). In the case where all marginal component and centering distributions form conjugate models, that is when $H_l(\theta_l)$ is the conjugate prior for $g_l(\cdot|\theta_l)$, then Gibbs sampling methods for conjugate models may be used. For the more general case, we recommend the Auxiliary Gibbs sampler of Neal (2000) for full conditional cluster updates. The Auxiliary Gibbs sampler is one of the most flexible alternatives for obtaining clustering conformations in that it requires the ability to sample directly from $H_l(\theta_l)$, but not that it be a conjugate prior. We use the Auxiliary Gibbs sampler with one auxiliary component for all of the analysis in this paper. Given clustering information, it is possible to update θ_j independently of all θ_l, $l \neq j$. Therefore, within a cluster it is possible to treat θ_j according to a parametric Bayes model with the centering distribution component H_j treated as the prior for θ_j, and only considering data $\boldsymbol{x}_{ip_j}$ for which the ith observation is included in the current cluster. Both Metropolis and Gibbs updating schemes are suitable for this purpose. A synopsis of our general posterior sampling method is given in Figure 16. For an example of a sampling algorithm with specific component marginals, refer to the material on the nonparametric prior in Section 3.4.1.

4.3 Simulation Studies

4.3.1 Detecting Clustering and Correlation

To evaluate the behavior of DPD style models, we consider a test case with mixtures of bivariate normal distributions. Two test distributions were considered; the first is a single

1. Initialize the parameter values:

   (a) Choose an initial clustering. Two obvious choices are: (1) one cluster containing all observations, or (2) each observation in a cluster by itself.

   (b) For each initial cluster of observations pairs, initialize the value of each marginal parameter set by sampling from the appropriate marginal centering distribution.

2. Obtain draws from the posterior distribution by repeating the following:

   (a) Given all parameter values, update the clustering configuration using one scan of the Auxiliary Gibbs sampler of Neal (2000).

   (b) For each component distribution, update the relevant parameter set θ_j for each cluster.

   (c) To obtain a sample from the posterior distribution of $\boldsymbol{\theta}$, select an existing cluster with probability proportional to the number of members, or a new cluster with probability proportional to τ_0. If an existing cluster is chosen, take the parameters from that cluster. If a new cluster is chosen, draw a parameter set from the centering distribution.

Figure 16: General computational procedure for DPDM model fitting and sampling.

bivariate normal distribution, and the second is a mixture of two bivariate normal distributions. They are described in Table 3.

A DPD model was developed which treated each cluster as consisting of the product of independent normal distributions $f_1(x_{i1}|\mu_{i1}, \sigma^2_{i1})f_2(x_{i2}|\mu_{i2}, \sigma^2_{i2})$. All association between the two random variables comes from the clustering induced by the Dirichlet process. This model is philosophically similar to the multivariate kernel density estimator of Epanechnikov (1969), for which each dimension had only a single bandwidth parameter.

Model fits were generated for samples of size 50, 100, and 500 for each test distribution. For each test distribution and sample size, two MCMC chains were run for 6,000 iterations with the first 1,000 discarded as burn in, and 1-in-10 thinning. This gave 1,000 draws from the posterior distribution. The standard Gibbs sampler for univariate normal data was used for updates. The centering distributions H_1 and H_2 both took $1/\sigma^2_i \sim \chi^2_1(1)$

Table 3: Distributions used for simulation studies. The first is a single bivariate normal distribution with negative covariance, while the second is a mixture of two bivariate normal distributions.

Test Distribution 1

weight	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}
1.0	0.0	0.0	0.25	0.5	-0.25

Test Distribution 2

weight	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}
0.5	1.0	-1.0	0.25	0.25	-0.20
0.5	-1.0	1.0	0.25	0.25	0.10

and $\mu_i \sim N(0, \sigma_i^2)$. The DP mass parameter was set to 1. Our two MCMC chains differed only in whether all observations started in a single cluster or were initialized in n distinct clusters. The resulting density estimates are displayed in Figure 17.

The DPD model quickly identifies the distinct component distributions for the mixture case, but takes longer to pick up on the within component correlation. This is unsurprising, as the only way to model this correlation is with overlapping clusters. These results indicate that the DPDM performs best when the prior assumption of independence given clustering holds.

4.3.2 Effect of the Mass Parameter

Most prior parameters for a DPD model will depend on what component marginal distributions are being used. However, the mass parameter τ_0 is common to all DPD models, and so warrants special attention in this general discussion. The role of τ_0 in mixture modeling is described by Ferguson (1983). He shows that as $\tau_0 \to 0$, it becomes increasingly likely that all observations will be clustered together. This extreme case results in a Bayesian parametric density estimate. On the other hand, as $\tau_0 \to \infty$, it becomes increasingly likely that all observations will belong to distinct clusters. The resulting density estimate will be imprecise, and not terribly useful. However, in the space between the extrema there is a

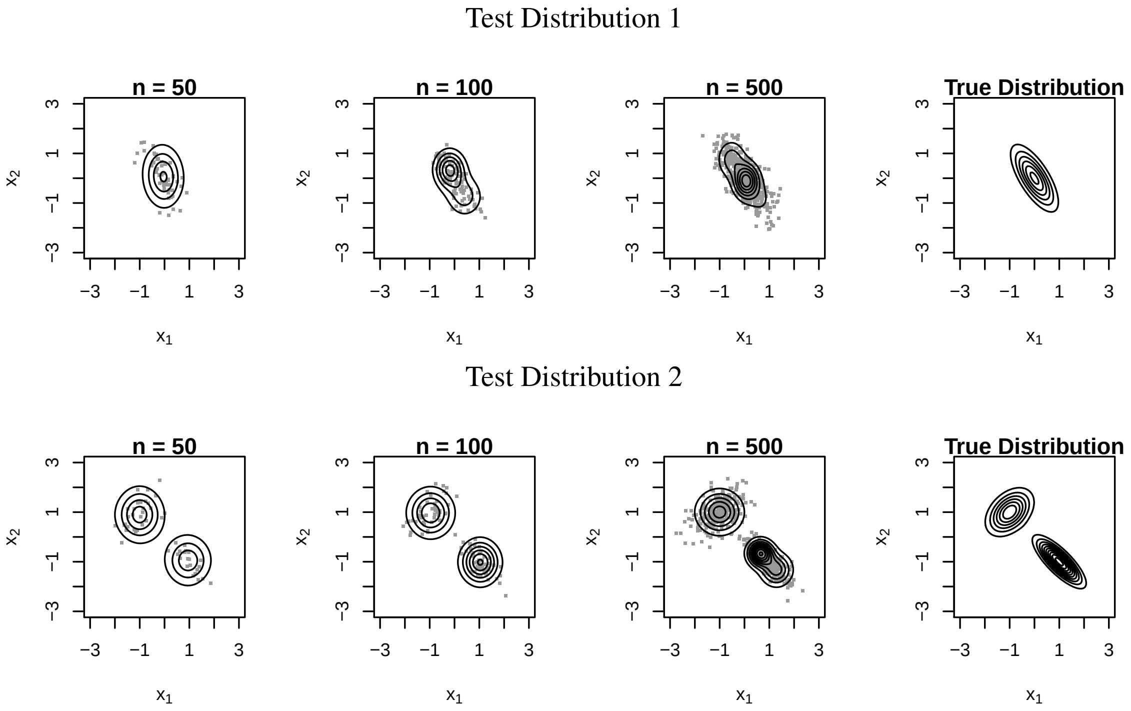


Figure 17: Density plots for DPD model. Each column is a different sample size, while each row is a separate test distribution.

considerable amount of interesting behavior to explore for τ_0.

Since increasing τ_0 boosts our expected number of clusters, the manipulation of this parameter could be useful for DPD models. Specifically, since all association between components is modeled via clustering, increasing the number of clusters should increase the sensitivity to dependence. However, the benefits of increasing the number of clusters must be weighed against the disadvantage of decreasing the number of observations in each cluster.

To test this theory, we again considered sample sizes $n = 50$, 100, and 500, as well as mass parameter values $\tau_0 = 0.1$, 1, 3, 6 and 10. We obtained 100 sample sets for each sample size n from our single normal test distribution from the previous section. Recall that this distribution only displayed within component correlation. We generate a density estimate using the same methods described in the previous section for each sample set and

Table 4: Summary of average divergence scores for various sample size and mass parameter settings.

	τ_0				
	0.1	1	3	6	10
$n = 50$	27.178	25.785	25.494	26.010	27.364
$n = 100$	16.375	15.319	15.303	15.760	16.651
$n = 500$	4.441	4.297	4.333	4.408	4.563

τ_0 value.

To gauge the similarity of our density estimates to the true distribution, we employed the Jensen-Shannon divergence on a 100×100 grid over the region $[-3, 3] \times [-3, 3]$. The formula for this divergence score is:

$$\frac{1}{2}\left(D_{KL}\left(\hat{f}, \frac{\hat{f}+f}{2}\right) + D_{KL}\left(f, \frac{\hat{f}+f}{2}\right)\right)$$

where D_{KL} is the Kullback-Leibler divergence defined by $D_{KL}(f, \hat{f}) = \sum_i f(i)\log\left(f(i)/\hat{f}(i)\right)$, f is our true distribution function, and $\hat{f}$ is our density estimate. The lower the divergence score, the closer our density estimate is to the true distribution. The average of the 100 divergence scores for each n, τ_0 combination is given in Table 4.

As we expect, divergence drops as sample size increases. The influence of the mass parameter τ_0 fades as sample size increases, although the optimal value changes little, being 3 for $n = 50$, 100 and 1 for $n = 500$. It is somewhat disappointing to note that adjusting the mass parameter alone does not offer dramatic improvements in the density estimate. As previously mentioned, increasing the mass parameter increases the expected number of clusters and thus reduces the expected cluster size. This is probably the cause of the degradation of model quality seen for very large values of the mass parameter.

4.3.3 Marginal Distributions

Since one of the attractive qualities of a DPD model is the ability to specify particular marginal distribution structures, it would be interesting to see how this model fares against

Table 5: Summary of average divergence scores for marginal models. Notice that the purely marginal models outperform the DPDM based marginals.

Marginal Models for x_1

				τ_0		
		0.1	1	3	6	10
Marginal Model	$n = 50$	0.104	0.143	0.240	0.379	0.563
	$n = 100$	0.049	0.070	0.116	0.186	0.277
	$n = 500$	0.008	0.013	0.021	0.036	0.053
DPDM Marginal	$n = 50$	0.106	0.146	0.215	0.310	0.431
	$n = 100$	0.089	0.103	0.129	0.171	0.230
	$n = 500$	0.027	0.028	0.033	0.039	0.048

Marginal Models for x_2

				τ_0		
		0.1	1	3	6	10
Marginal Model	$n = 50$	0.083	0.105	0.164	0.258	0.380
	$n = 100$	0.042	0.054	0.084	0.126	0.189
	$n = 500$	0.008	0.010	0.015	0.024	0.036
DPDM Marginal	$n = 50$	0.084	0.098	0.125	0.172	0.232
	$n = 100$	0.070	0.076	0.085	0.101	0.127
	$n = 500$	0.031	0.029	0.030	0.030	0.032

the marginal model alone. That is, we wish to know how the marginal distribution derived from a DPD model fit on a multivariate dataset compares to the marginal distribution fit on the variable of interest alone. We fit such marginal models on the 300 samples discussed in Section 4.3.2. We used the same scheme, including number of MCMC chains and prior parameter settings, described in the previous section with the only difference being that we were now fitting a univariate instead of bivariate distribution. We calculated the Jensen-Shannon divergence on 100 equally spaced points between -3 and 3 for both our new marginal models and the marginal distribution estimates derived from our bivariate DPDM density estimate as described in Section 4.2.2. The results are given in Table 5.

Notice how marginal models generally perform best for $\tau_0 = 0.1$. This makes sense, as our test distribution has univariate normal margins. However, it means that the DPDM generally has different optimal mass parameters for joint and marginal modeling. Notice how the best marginal model outperforms the best DPDM marginal even for the larger sample sizes. This could be attributed to the fact that the DPDM is designed for joint modeling, and so loses some efficiency in the marginals by generating clusters based on multiple random variables simultaneously. These results also suggest that marginal and DPDM models have different optimal mass parameter settings, and this fact should be taken into account when choosing between modeling data with independent marginals or a DPDM. The fact that the DPDM is at a disadvantage modeling marginal distributions also indicates that in cases with weak dependence, the DPDM may be outperformed by a model based on independent marginals. This situation will arise when the advantages of dependence modeling are overwhelmed by the degradation of performance in the marginal distributions. These factors will come into play, and must be accounted for, when we apply our DPDM model to real data with an unknown joint distribution..

4.4 Application to Protein Relative Packing Groups

4.4.1 Introduction to Protein Cliques

A protein is a long, unbranched chain of amino acids. These chains fold up into three-dimensional conformations which are of keen interest to biologists, as a protein's structure determines its function and behavior in biological systems. There are 20 naturally occurring amino acids, each of which in composed of four heavy backbone atoms (an alpha carbon atom C_α, a carbonyl carbon atom C, an oxygen atom O, and a nitrogen atom N) and a distinctive side-chain. While the backbone atoms are common to all amino acids, the side-chains are all different and in large part determine the final structure of the protein.

A common simplified representation of the protein backbone is a sequence of (ϕ, ψ) angle pairs, one for each sequence position. This representation was initially proposed by Ramachandran et al. (1963). We will also simplify the side-chain representation by using the Cartesian coordinates of the side-chain centroid, which we define as the location of the center of mass of the side-chain relative to the C_α atom of the amino acid backbone.

An outstanding problem in structural biology is modeling the relationship between backbone and side-chain conformations for protein relative packing groups (RPGs), first defined by Holmes and Tsai (2005). A protein clique is a set of amino acids which are in close contact when a protein is folded. A relative packing group is a set of protein cliques with certain shared characteristics. An RPG may be composed of cliques from members of a single protein family, or from similarly packed regions across a variety of different protein types.

Since relative packing groups are a recent development in structural biology, the relationship between amino acid positions and backbone conformation for members of cliques has not been studied. The distributions have been characterized separately (Day et al., 2010), but joint modeling would provide additional information about typical clique behavior within an RPG. Joint distributions for the behavior of torsion angles and side-chain centroids would also be invaluable in the area of structure prediction. Such distributions could be used for applications such as generating candidate backbone-centroid conformations for proposed structures, or for evaluating the feasibility of structures developed by alternative methods. Note that these applications would require rapid conditional and unconditional sampling and density evaluation respectively.

The necessity for mixture models has precluded the use of existing linear-angular models (Johnson and Wehrly, 1978) due to computational intractability. However, the existence of relatively straightforward marginal mixture models makes this problem an excellent candidate for a DPDM. In the remainder of this section we will present such a model, and use it to study a relative packing group from immunoglobulin-binding proteins.

4.4.2 Methods

We propose a Dirichlet process dependence model for association between protein torsion angles and centroid coordinates. Our model can be summarized as follows:

$$\begin{aligned}
\phi_i, \psi_i \mid \mu_i, \nu_i, \Omega_i &\sim g_1(\phi_i, \psi_i \,|\mu_i, \nu_i, \Omega_i) \\
x_i, y_i, z_i \mid \boldsymbol{\theta}_i, \Sigma_i^{-1} &\sim g_2(x_i, y_i, z_i \,|\boldsymbol{\theta}_i, \Sigma_i^{-1}) \\
\mu_i, \nu_i, \Omega_i, \boldsymbol{\theta}_i, \Sigma_i^{-1} | \, G &\sim G \\
G &\sim DP(\tau_0 H_1 H_2).
\end{aligned}$$

Here, the distribution G is a draw from a Dirichlet process with centering distributions H_1 and H_2 for the parameter sets (μ, ν, Ω) and $(\boldsymbol{\theta}, \Sigma^{-1})$ respectively. The distribution g_1 is a bivariate von Mises sine model with mean parameters μ, ν, and precision matrix Ω as described in previous chapters. The distribution g_2 is a trivariate normal distribution with mean vector $\boldsymbol{\theta}$ and precision matrix Σ^{-1}. The centering distribution H_1 is the product of a bivariate von Mises sine model and a Wishart distribution, and sampling can proceed according to the algorithm described in Chapter II. H_2 is the normal-Wishart conjugate prior distribution for multivariate normal data with unknown mean vector and covariance matrix.

Unless otherwise noted, all calculations were carried out with the following prior parameter settings. The mass parameter τ_0 was equal to 3. The bivariate von Mises component of H_1 had a mean of $(0, 0)$ and the precision matrix equal to $0.1I_2$ where I_n is the $n \times n$ identity matrix. The Wishart component had a shape parameter $\alpha = 1$ and scale matrix $\beta = 0.1I_2$. The multivariate normal-Wishart H_2 had Wishart parameters $\alpha = 1.5$ and $\beta = 0.25I_3$, while the normal component had a mean of 0 and a scaling factor $\lambda = 1$. Note that all parameterizations are consistent with those of Bernardo and Smith (1994), and thus the Wishart distributions have an expected value of $\alpha\beta^{-1}$. Most parameters were chosen to minimize the influence of the prior distribution, or because they were found to work well in practice. The Cartesian coordinate data points for each clique position were centered at the origin before modeling.

For each model, two MCMC chains were run for 6,000 iterations, the first 1,000 of which served as burnin. Using 1-in-10 thinning, this gave 1,000 draws from the posterior distribution of the parameter sets. One chain was initialized with all observations in a single cluster, while the other started with all observations in individual clusters.

4.4.3 Testing for Dependence

We consider a dataset based on a highly populated relative packing group which is characterized by cliques of four residues with one residue in a region of protein with helical structure and the remaining three residues coming from two different strand structure regions. We will consider a set of 61 such cliques from the immunoglobulin-binding protein G and protein L domains. These are proteins produced by bacteria which bind to human antibodies. Plots of angle pairs and centroid locations for each position are shown in Figure 18. Angle pairs for positions 1,3, and 4 show typical strand conformations, while the torsion angles for position 2 are in a helical region.

One potential avenue for biological investigation is to determine whether or not clique side-chain locations constrain the conformation of a protein's backbone. This is equivalent to the existence of dependence between side-chain centroid and torsion angle distributions. We designed a permutation test to determine if there was identifiable dependence between torsion angles and centroid positions. For each clique position we randomly divided the data into six groups: five of ten and one of eleven. We then generated permutation density estimates in the following manner: we took all data excluding a single group, and considered this to be our training set. We generated one hundred permutations of the centroid coordinates for each set, thus ensuring that angle pairs and centroid coordinates were

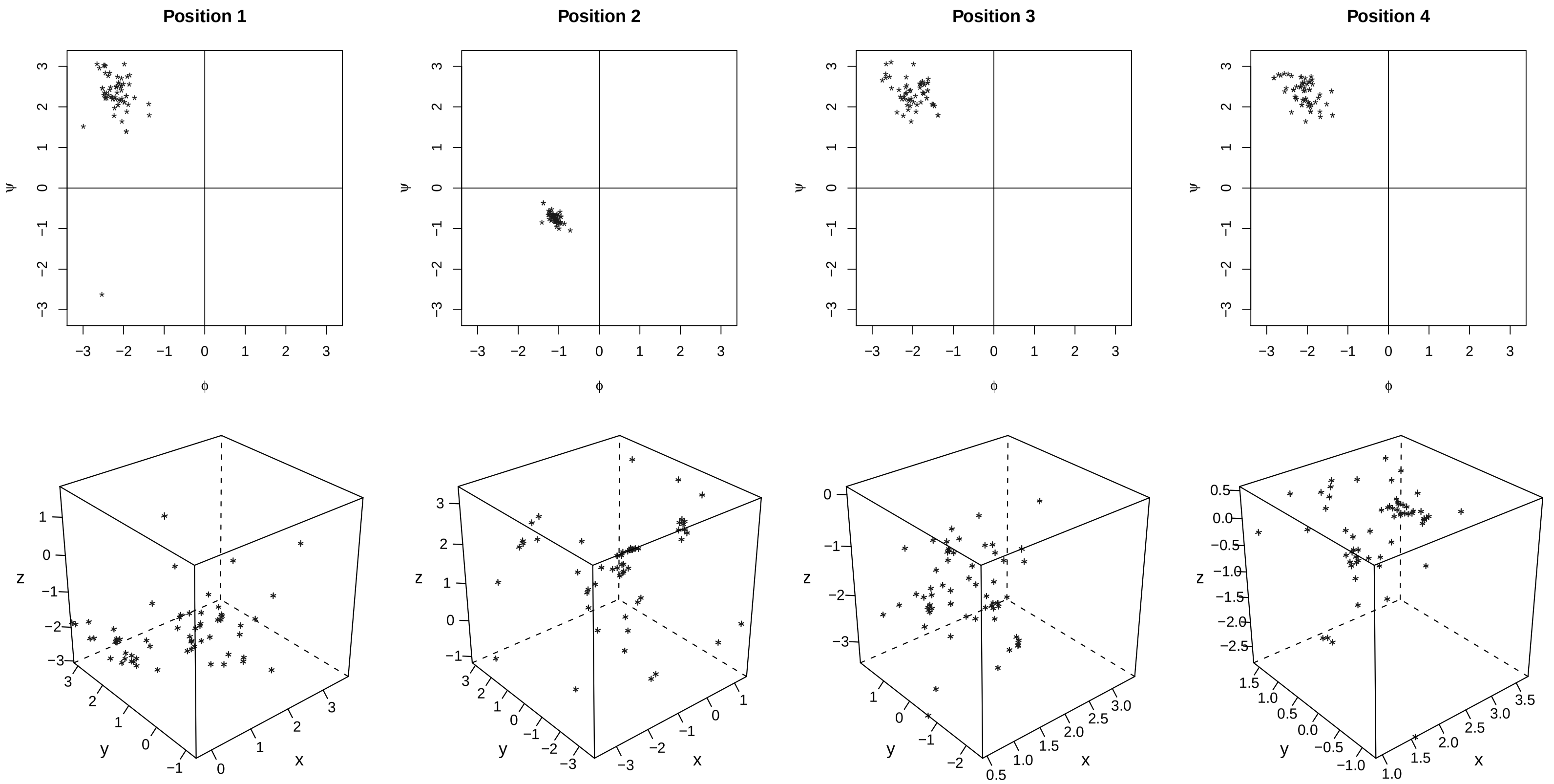


Figure 18: Data for the four positions of a relative packing group for immunoglobulin-binding protein G and protein L domains. The top plots show (ϕ, ψ) torsion angle pairs while the lower plots show (x, y, z) centroid coordinates.

matched randomly. (Note that individual angle pairs were not permuted, nor were individual centroid coordinates.) This generated 100 datasets with the same marginal structure as the true training data, but with potential dependence between angle pairs and coordinates removed. By comparing density estimates based on this altered data to those based on the original, we can determine whether or not incorporating dependence structure is important for modeling the observed data.

Using the model described in Section 4.4.2, we generated DPDM density estimates on each of the permuted training sets as well as for the original data. This gave us a total of 606 density estimates per clique position. We then generated an estimated density statistic for each permutation. We define the estimated density value to be:

$$V_i = \prod_{j=1}^{n} \hat{f}_{i,-j}(\phi_j, \psi_j, x_j, y_j, z_j)$$

where i denotes the ith permutation and the subscript $-j$ indicates that the density estimate $\hat{f}$ was fit on a dataset not including the jth observation. Note that, due to the fact that permutation occurs after the division of the data into groups, the assignment of six given permutation groups together into a single cross-validation set i is arbitrary. We also calculated this statistic for density estimates fit on the unpermuted dataset. The results are summarized in Figure 19. The boxplots show the distribution of $\log_{10}(V_i)$ for the 100 permutations for each sequence position, while stars mark the value of the $\log_{10}$ estimated density for the true dataset.

Clique positions 1, 3, and 4 all show strong evidence of dependence, with the latter two positions being particularly striking. In all of these cases the test statistic for density estimates based on unpermuted data showed a higher probability than any of the 100 permuted data density estimates. There was no evidence of association at position 2, with the unpermuted data statistic in the 37th percentile of the permutation distribution.

These results are interesting in that they demonstrate that, while dependence certainly

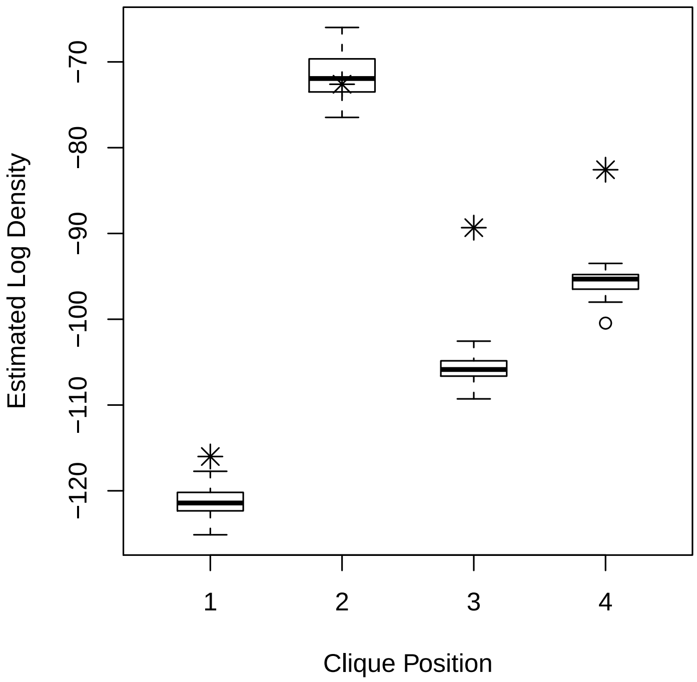


Figure 19: Summary of the permutation test for clique data. Boxplots show the estimated log densities for permuted data, while the stars show the value for the true dataset.

may exist between a side-chain centroid and the protein backbone, there exist certain clique positions where it either does not exist or can not be effectively modeled using our technique. In particular, while we detected dependence at all three sequence positions in sheet regions, none was found at the single position in a helical region. To further explore this situation we compared the density estimates generated by the DPD model and an appropriate independence model. Specifically, we fit the DP mixture of bivariate von Mises distributions from Chapter II for the torsion angles and a separate DP mixture of normals for the centroid coordinates.

In order to compensate for the differing preferences for mass parameter values between the joint and marginal models, we fit density estimates using τ_0 equal to 0.1, 1, 3,

Table 6: Summary of comparison between independence and DPD models. Numbers in the density column are the $\log_{10}$ of the estimated density statistic. The value for τ_0 corresponds to the mass parameter value which gave the highest estimated density. Note that the independence model is the product of the angle and centroid models.

	Clique Position							
	1		2		3		4	
	Density	τ_0	Density	τ_0	Density	τ_0	Density	τ_0
Angles	-18.204	1	31.522	6	-18.077	0.1	-13.156	1
Centroid	-91.181	1	-95.434	1	-77.346	3	-72.346	3
Independence Model	-109.385	-	-63.913	-	-95.423	-	-85.502	-
DPD Model	-115.910	6	-72.607	3	-89.326	3	-82.571	3

6, and 10. Using the formula derived by Antoniak (1974), we find that this gives prior expected numbers of clusters of about 1.4, 5, 9, 14, and 19 respectively. We generate model fits and use the estimated density in the same manner as before. The results are shown in Table 6, which gives the highest estimated log density value and the τ_0 for which it was achieved. Note that for the independent angle and centroid models the optimal τ_0 is chosen separately. The independence model refers to the product of the best angle and centroid models for a given clique position.

We can treat our estimated densities as components of a Bayes factor with no prior preference between dependence and independence by taking the difference between the log densities for the DPD and independence models. We find that the differences are -6.525, -8.964, 6.097, and 2.931. All of these values represent substantial evidence for or against the DPDM model according to the criteria of Kass and Raftery (1995). The dependence model for position 2 is the most strongly disfavored, which agrees with our permutation test result which found no dependence there. Interestingly the DPDM model is also disfavored at position 1, for which we did identify dependence previously. This is probably due to the fact we are dealing with moderate sample sizes, and the DPDM marginal model penalty could be overwhelming the advantages of identifying dependence at this position. This result highlights the fact that a straight comparison between a marginal independence and DPDM model is not the best way to test for dependence, as factors such as the marginal modeling penalty and different optimal prior parameter settings could lead to misleading results. Positions 3 and 4 are substantially better represented by the DPDM model than the marginal independence model. Recalling the permutation test results, these positions were also the most dramatically different from the permutation distribution.

The results are interesting from a biological perspective as there is an apparent relationship between secondary structure type and dependence between the side-chain location and torsion angles. The strongest evidence of dependence was found for clique positions in strand secondary structure regions, while none was found for the helical position. In regards to the results in the second portion of our analysis, it is interesting to note that clique positions 3 and 4 are actually local to one another on the backbone. This suggests that the region of the backbone where these clique positions are located is more constrained than the region corresponding to position 1. A comparatively small change in centroid location, which might not influence position 1, has the potential to noticeably rearrange the backbone at positions 3 or 4.

The results also serve to highlight a few interesting statistical points. First, we see that the dependence model tends to favor higher τ_0 values than the marginal models, although the pattern does not always hold within a clique position. Secondly, we see the disadvantage of using a DPDM where there is no apparent dependence. The sacrifice in the quality of marginals suffered by the DPD model at clique position 2 puts it at a severe disadvantage when compared to the independence model. Similar results at position 1 suggest that dependence at that location, while identifiable, is weak. For this reason we recommend that the DPD model should be used with some care, particularly for small to moderate sample sizes.

4.5 Discussion

We have proposed a new method for nonparametric modeling of dependence incorporating Dirichlet process mixtures of component distributions. Our method provides advantages over nonparametric copula models in its natural handling of multivariate marginal distributions and straightforward conditional computation.

In applying our model to protein structure data we were able to demonstrate the value of joint modeling between side-chain position and backbone conformation. In the course of this analysis, we presented a DPDM permutation test for dependence. We discovered that the level dependence will vary from position to position in a protein RPG. This information can be used to increase the efficiency of protein structure prediction models by incorporating dependence structure only where it is needed.

A brief mention is warranted for some potential issues regarding the efficiency of our proposed MCMC scheme. The results in Chapter III indicated that the noninformative prior model, which followed our proposed DPDM framework, experienced mixing problems as the number of component distributions increased. This suggests that as the number of component distributions increases, vigilance for MCMC issues should also increase. While there is no theoretical limit to the number of component marginals, this issue will impose practical restrictions.

CHAPTER V

CONCLUSIONS

We have presented a number of innovative statistical models designed to answer open questions in the protein structure prediction community. The material from Chapter II allowed density estimation to be performed for torsion angle data with much smaller sample sizes than had been possible with existing binning methods. This let us describe the behavior of the so-called half positions for torsion angle data in a rigorous statistical way. This analysis demonstrated that half position distributions are better suited than whole positions for template-based modeling.

Our extension of this single position model in Chapter III permitted the application of template-based modeling methods to loop regions for the first time. In addition to a straightforward nonparametric model, we also took full advantage of the ability of Bayesian density estimation to incorporate prior information by presenting the DPM-HMM. This allowed us to leverage information from surrounding sequence positions to obtain informative density estimates even at alignment positions with few or no observed data points. We found our method to be an effective means of compensating for the sparse data problem in protein loop regions, and our method proved superior to existing alternatives for loop modeling.

Our final contribution to the field of protein structure prediction was the development of a model which would link together the backbone and side-chain conformations. This involved a multivariate angular-linear data mix, and in our nonparametric context existing joint models were unsuitable. We therefore defined a class of Dirichlet process dependence models in Chapter IV which have attractive properties in terms of modeling flexibility and computation. We applied this method to a clique dataset to investigate the joint behavior of torsion angle pairs and side-chain centroids. We arrived at the somewhat surprising result that while joint modeling is often desirable for this data, this is not always the case. This insight opens up new areas of inquiry in the study of protein cliques, and also allows for more efficient structure prediction. By presenting methods which allow us to test the efficacy of DPDM models, we can pursue improved prediction strategies which incorporate dependence only where there is strong evidence that it exists.

One of the hallmarks of these research projects was the mutual advancement of statistics and biology through the development of new statistical methods. Each biological problem under study was either not addressed or insufficiently addressed by existing statistical methodology, and improved methodology suggested areas of biological interest which were not accessible previously. The purpose of all of this statistical development was to improve template-based structure prediction strategies to be employed in the CASP (Critical Assessment of Techniques for Protein Structure, see e.g. Moult, 2005) 9 experiment. CASP is an international structure prediction competition, and the methods presented in this dissertation are being employed by a CASP competition team. The DPM-HMM method from Chapter III is being used to generate candidate loops for a protein structure prediction algorithm, while the DPDM for torsion angles and side-chain centroids from Chapter IV is being employed as a scoring function for side-chain placement. Competitions such as this one provide an excellent avenue for the application of new statistical ideas, and the introduction of advanced statistical modeling to the wider scientific community.

Beyond CASP, the general statistical strategies developed are not limited to biological applications. The methods presented in Chapters II and III are suitable for other bivariate or multivariate angular data situations, such as characterizing wind direction. The idea of combining an informative prior with a Bayesian nonparametric technique, as in the case of the DPM-HMM, can be applied in any density estimation setting, and presents an advantage over frequentist kernel density techniques. The DPD model provides an alternative to nonparametric copulas wherever they might be applied, and offers key advantages particularly in terms of modeling multivariate marginals. The further applicability of these strategies, all initially developed to solve structural biology problems, emphasizes the versatility of the field of statistics.

REFERENCES

Anderson, A. G. and Hermans, J. (1988). Microfolding: Conformational probability map for the alanine dipeptide in water from molecular dynamics simulations. *Proteins* **3,** 262–265.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2,** 1152–1174.

Bagchi, P. and Guttman, I. (1988). Theoretical considerations of the multivariate von Mises-Fisher distribution. *Journal of Applied Statistics* **15,** 149–169.

Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science* **294,** 93–96.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Advances in Neural Information Processing Systems 14*, Dietterich, T., Becker, S., and Ghahramani, Z., (eds.), 577–584. Cambridge, Massachusetts, MIT Press.

Berman, H. M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural Biology* **10,** 980.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester, West Sussex, England, John Wiley & Sons.

Berry, D. A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *The Annals of Statistics* **7,** 558–568.

Bonneau, R. and Baker, D. (2001). Ab initio protein structure prediction: Progress and prospects. *Annual Review of Biophysics and Biomolecular Structure* **30,** 173–189.

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proceedings of the National Acadademy of Sciences* **105,** 8932–8937.

Butterfoss, G. L., Richardson, J. S., and Hermans, J. (2005). Protein imperfections: Separating intrinsic from extrinsic variation of torsion angles. *Acta Crystallographica, Section D: Biological Crystallography* **61,** 88–98.

Chen, S. X. and Huang, T. (2005). Nonparametric estimation of copula functions for dependence modeling. *Canadian Journal of Statistics* **35,** 265–282.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75,** 79–97.

Clore, G. M. and Schwieters, C. D. (2002). Theoretical and computational advances in biomolecular NMR spectroscopy. *Current Opinions in Structural Biology* **12,** 146–153.

Dahl, D. B., Bohannan, Z., Mo, Q., Vannucci, M., and Tsai, J. W. (2008). Assessing side-chain perturbations of the protein backbone: A knowledge based classification of residue Ramachandran space. *Journal of Molecular Biology* **378,** 749–758.

Das, R. and Baker, D. (2008). Macromolecular modeling with Rosetta. *Annual Review of Biochemistry* **77,** 363–382.

Day, R., Lennox, K. P., Dahl, D. B., Vannucci, M., and Tsai, J. W. (2010). *Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure*. Submitted.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99,** 205–215.

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A. (2007). The protein folding problem: When will it be solved? *Current Opinion in Structural Biology* **17,** 342–346.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **69,** 163–183.

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104,** 1042–1051.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications* **14,** 153–158.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1,** 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, Rizvi, H. and Rustagi, J., (eds.), 287–302. New York, New York, Academic Press.

Fitzkee, N. C., Fleming, P. J., and Rose, G. D. (2005). The Protein Coil Library: A structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins* **58,** 852–854.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100,** 1021–1035.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82,** 543–552.

Grail, B. M. and Payne, J. W. (2000). Predominant torsional forms adopted by dipeptide conformers in solution: Parameters for molecular recognition. *Journal of Peptide Science* **6,** 186–199.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandanavian Journal of Statistics* **28,** 355–375.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* **101,** 179–194.

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

Ho, B. K., Thomas, A., and Brasseur, R. (2003). Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Science* **12,** 2508–2522.

Holmes, J. B. and Tsai, J. W. (2005). Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. *Journal of Molecular Biology* 706–721.

Hovmoller, S., Zhou, T., and Ohlson, T. (2002). Conformations of amino acids in proteins. *Acta Crystallographica, Section D: Biological Crystallography* **58,** 768–776.

Johnson, R. A. and Wehrly, T. E. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association* **73,** 602–606.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., Sander, C., and England, E. (1997). Predicting protein structure using hidden Markov models. *Proteins: Structure, Function, and Genetics* **29,** 134–139.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Kopp, J., Bordoli, L., Battey, J. N. D., Kiefer, F., and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69(Suppl 8),** 38–56.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P. E., and Berman, H. M. (2006). The RCSB PDB information portal for structural genomics. *Nucleic Acids Research* **34,** D302–305.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM Journal on Scientific and Statistical Computing* **7,** 60–71.

Lee, S. Y. and Skolnick, J. (2008). Benchmarking of TASER 2.0: An improved protein structure prediction algorithm with more accurate predicted contact restraints. *Biophysical Journal* **95,** 1956–1964.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12,** 351–357.

Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003). Structure validation by Calpha geometry: Phi, psi and Cbeta deviation. *Proteins* **50,** 437–450.

MacEachern, S. N. (2000). *Dependent Dirichlet Processes*. Technical report, Ohio State University, Department of Statistics.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7,** 223–238.

Mardia, K. V. (1975). Statistics of directional data (Com: P371-392). *Journal of the Royal Statistical Society, Series B: Methodological* **37,** 349–371.

Mardia, K. V. (2009). Statistical complexity in protein bioinformatics. In *LASR 2009 Proceedings: Statistical Tools for Challenges in Bioinformatics*, Gustano, A., Mardia, K. V., and Fallaize, C. J., (eds.), 9–20. Leeds, Leeds University Press.

Mardia, K. V. and El-Atoum, S. A. M. (1976). Bayesian inference for the von Mises-Fisher distribution. *Biometrika* **63,** 203–205.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics* **36,** 99–109.

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63,** 505–512.

McGuffin, L. J., K., B., and T., J. D. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16,** 404–405.

Michalsky, E., Goede, A., and Preissner, R. (2003). Loops In Proteins (LIP) - a comprehensive loop database for homology modeling. *Protein Engineering Design & Selection* **16,** 979–985.

Moult, J. (2005). A decade of CASP: Progress, bottlenecks, and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15,** 285–289.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19,** 95–110.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York, New York, USA, Springer.

Osguthorpe, D. J. (2000). Ab initio protein folding. *Current Opinion in Structural Biology* **10,** 146–152.

Pertsemlidis, A., Zelinka, J., Fondon, J. W., Henderson, R. K., and Otwinowski, Z. (2005). Bayesian statistical studies of the Ramachandran distribution. *Statistical Applications in Genetics and Molecular Biology* **4,** Article 35.

Pewsey, A. and Jones, M. (2005). Discrimination between the von Mises and wrapped normal distributions: Just how big does the sample size have to be? *Statistics* **39,** 81–89.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Molecular Biology* **7,** 95–99.

Rivest, L. P. (1982). Some statistical methods for bivariate circular data. *Journal of the Royal Statistical Society, Series B: Methodological* **44,** 81–90.

Rivest, L. P. (1988). A distribution for dependent unit vectors. *Communications in Statistics: Theory and Methods* **17,** 461–483.

Rodrigues, J., Leite, J. G. a., and Milan, L. A. (2000). An empirical Bayes inference for the von Mises distribution. *Australian & New Zealand Journal of Statistics* **42,** 433–440.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103,** 1131–1144.

Rother, D., Sapiro, G., and Pande, V. (2008). Statistical characterization of protein ensembles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **45,** 42–55.

Schlick, T. (2006). *Molecular Modeling and Simulation: An Interdisciplinary Guide*. New York, New York, USA, Springer.

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association* **97,** 337–351.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4,** 639–650.

Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89,** 719–723.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101,** 1566–1581.

Tiwari, R. C., Jammalamadaka, S. R., and Chib, S. (1988). Bayes prediction density and regression estimation–a semiparametric approach. *Empirical Economics* **13,** 209–222.

Usón, I. and Sheldrick, G. M. (1999). Advances in direct methods for protein crystallography. *Current Opinion in Structural Biology* **9,** 643 – 648.

West, M. (1990). *Bayesian Kernel Density Estimation*. Discussion Paper 90-A02, Duke University, Institute of Statistics and Decision Sciences.

Xing, E. P. and Sohn, K. A. (2007). Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis* **2,** 501–528.

Xue, B., Dor, O., Faraggi, E., and Zhou, Y. (2008). Real-value prediction of backbone torsion angles. *Proteins* **72,** 427–433.

APPENDIX A

DERIVATION OF FULL CONDITIONAL DISTRIBUTION IN CHAPTER II

We will consider a general eight parameter bivariate von Mises distribution. Using the representation from Mardia et al. (2007), the density can be expressed as:

$$\begin{aligned} f(\phi_i, \psi_i) \propto \exp\{&\kappa_{1i}\cos(\phi_i - \mu) + \kappa_{2i}\cos(\psi_i - \nu) + \\ &[\cos(\phi_i - \mu), \sin(\phi_i - \mu)]A_i[\cos(\psi_i - \nu), \sin(\psi_i - \nu)]^T\} \end{aligned}$$

where A_i is a 2×2 matrix. For a dataset consisting of (ϕ_i, ψ_i), $i = 1, ...n$, the full conditional log density of (μ, ν) up to a constant can be expressed as:

$$\begin{aligned} L(\mu, \nu) =& \sum_{i=1}^{n} \kappa_{1i}\cos(\phi_i - \mu) + \kappa_{2i}\cos(\psi_i - \nu) \\ &+ [\cos(\phi_i - \mu), \sin(\phi_i - \mu)]A_i[\cos(\psi_i - \nu), \sin(\psi_i - \nu)]^T \\ =& \left(\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right)[\cos(\mu), \sin(\mu)]^T \\ &+ \left(\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right)[\cos(\nu), \sin(\nu)]^T \\ &+ \sum_{i=1}^{n}[\cos(\phi_i - \mu), \sin(\phi_i - \mu)]A_i[\cos(\psi_i - \nu), \sin(\psi_i - \nu)]^T. \end{aligned}$$

Notice that the first two terms are consistent with a bivariate von Mises distribution with:

$$\tilde{\mu} = \arctan\left(\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right) \qquad \tilde{\nu} = \arctan\left(\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right)$$

$$\tilde{\kappa}_1 = \left|\sum_{i=1}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right| \qquad \tilde{\kappa}_2 = \left|\sum_{i=1}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right|.$$

The full conditional means are the directions of the sums of the observation vectors, while the full conditional concentration parameters are the magnitudes of the same sums. This allows us to rewrite the log likelihood as:

$$\begin{aligned}=&\tilde{\kappa}_1\cos(\mu-\tilde{\mu})+\tilde{\kappa}_2\cos(\nu-\tilde{\nu})\\&+\sum_{i=1}^{n}[\cos(\phi_i-\mu),\sin(\phi_i-\mu)]A_i[\cos(\psi_i-\nu),\sin(\psi_i-\nu)]^T.\end{aligned}$$

We will now focus on the final term of the log likelihood to determine $\tilde{A}$.

$$\begin{aligned}&\sum_{i=1}^{n}[\cos(\phi_i-\mu),\sin(\phi_i-\mu)]A_i[\cos(\psi_i-\nu),\sin(\psi_i-\nu)]^T\\
=&\sum_{i=1}^{n}[\cos(\mu),\sin(\mu)]\begin{bmatrix}\cos(\phi_i) & \sin(\phi_i)\\ \sin(\phi_i) & -\cos(\phi_i)\end{bmatrix}A_i\begin{bmatrix}\cos(\psi_i) & \sin(\psi_i)\\ \sin(\psi_i) & -\cos(\psi_i)\end{bmatrix}[\cos(\nu),\sin(\nu)]^T\\
=&[\cos(\mu),\sin(\mu)]\begin{bmatrix}\cos(\tilde{\mu}) & -\sin(\tilde{\mu})\\ \sin(\tilde{\mu}) & \cos(\tilde{\mu})\end{bmatrix}\begin{bmatrix}\cos(\tilde{\mu}) & -\sin(\tilde{\mu})\\ \sin(\tilde{\mu}) & \cos(\tilde{\mu})\end{bmatrix}^{-1}\\
&\left(\sum_{i=1}^{n}\begin{bmatrix}\cos(\phi_i) & \sin(\phi_i)\\ \sin(\phi_i) & -\cos(\phi_i)\end{bmatrix}A_i\begin{bmatrix}\cos(\psi_i) & \sin(\psi_i)\\ \sin(\psi_i) & -\cos(\psi_i)\end{bmatrix}\right)\\
&\begin{bmatrix}\cos(\tilde{\nu}) & \sin(\tilde{\nu})\\ -\sin(\tilde{\nu}) & \cos(\tilde{\nu})\end{bmatrix}^{-1}\begin{bmatrix}\cos(\tilde{\nu}) & \sin(\tilde{\nu})\\ -\sin(\tilde{\nu}) & \cos(\tilde{\nu})\end{bmatrix}[\cos(\nu),\sin(\nu)]^T\\
=&[\cos(\mu-\tilde{\mu}),\sin(\mu-\tilde{\mu})]\begin{bmatrix}\cos(\tilde{\mu}) & -\sin(\tilde{\mu})\\ \sin(\tilde{\mu}) & \cos(\tilde{\mu})\end{bmatrix}^{-1}\\
&\left(\sum_{i=1}^{n}\begin{bmatrix}\cos(\phi_i) & \sin(\phi_i)\\ \sin(\phi_i) & -\cos(\phi_i)\end{bmatrix}A_i\begin{bmatrix}\cos(\psi_i) & \sin(\psi_i)\\ \sin(\psi_i) & -\cos(\psi_i)\end{bmatrix}\right)\\
&\begin{bmatrix}\cos(\tilde{\nu}) & \sin(\tilde{\nu})\\ -\sin(\tilde{\nu}) & \cos(\tilde{\nu})\end{bmatrix}^{-1}[\cos(\nu-\tilde{\nu}),\sin(\nu-\tilde{\nu})]^T.\end{aligned}$$

Note that the determinants of the $\tilde{\mu}$ and $\tilde{\nu}$ matrices are both $\cos(0) = 1$.

$$
\begin{aligned}
=&[\cos(\mu - \tilde{\mu}), \sin(\mu - \tilde{\mu})] \left(\sum_{i=1}^{n} \begin{bmatrix} \cos(\tilde{\mu}) & \sin(\tilde{\mu}) \\ -\sin(\tilde{\mu}) & \cos(\tilde{\mu}) \end{bmatrix} \right. \\
&\begin{bmatrix} \cos(\phi_i) & \sin(\phi_i) \\ \sin(\phi_i) & -\cos(\phi_i) \end{bmatrix} A_i \begin{bmatrix} \cos(\psi_i) & \sin(\psi_i) \\ \sin(\psi_i) & -\cos(\psi_i) \end{bmatrix} \\
&\left. \begin{bmatrix} \cos(\tilde{\nu}) & -\sin(\tilde{\nu}) \\ \sin(\tilde{\nu}) & \cos(\tilde{\nu}) \end{bmatrix} \right) [\cos(\nu - \tilde{\nu}), \sin(\nu - \tilde{\nu})]^T \\
=&[\cos(\mu - \tilde{\mu}), \sin(\mu - \tilde{\mu})] \\
&\left(\sum_{i=1}^{n} \begin{bmatrix} \cos(\phi_i - \tilde{\mu}) & \sin(\phi_i - \tilde{\mu}) \\ \sin(\phi_i - \tilde{\mu}) & -\cos(\phi_i - \tilde{\mu}) \end{bmatrix} A_i \begin{bmatrix} \cos(\psi_i - \tilde{\nu}) & \sin(\psi_i - \tilde{\nu}) \\ \sin(\psi_i - \tilde{\nu}) & -\cos(\psi_i - \tilde{\nu}) \end{bmatrix} \right) \\
&[\cos(\nu - \tilde{\nu}), \sin(\nu - \tilde{\nu})]^T.
\end{aligned}
$$

So our full conditional matrix (with a uniform prior) will be:

$$
\tilde{A} = \sum_{i=1}^{n} \begin{bmatrix} \cos(\phi_i - \tilde{\mu}) & \sin(\phi_i - \tilde{\mu}) \\ \sin(\phi_i - \tilde{\mu}) & -\cos(\phi_i - \tilde{\mu}) \end{bmatrix} A_i \begin{bmatrix} \cos(\psi_i - \tilde{\nu}) & \sin(\psi_i - \tilde{\nu}) \\ \sin(\psi_i - \tilde{\nu}) & -\cos(\psi_i - \tilde{\nu}) \end{bmatrix}.
$$

Now consider the situation with a bivariate von Mises prior on (μ, ν) with parameters μ_0, ν_0, κ_{10}, κ_{20}, and $A_0 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. For the purposes of calculating $\tilde{\mu}$, $\tilde{\nu}$, $\tilde{\kappa}_1$, and $\tilde{\kappa}_2$ the prior can be treated as an additional observation with $\phi_0 = \mu_0$ and $\psi_0 = \nu_0$. The situation for the matrix $\tilde{A}$ is slightly more complicated. The full conditional matrix changes to:

$$
\begin{aligned}
\tilde{A} =& \left(\sum_{i=1}^{n} \begin{bmatrix} \cos(\phi_i - \tilde{\mu}) & \sin(\phi_i - \tilde{\mu}) \\ \sin(\phi_i - \tilde{\mu}) & -\cos(\phi_i - \tilde{\mu}) \end{bmatrix} A_i \begin{bmatrix} \cos(\psi_i - \tilde{\nu}) & \sin(\psi_i - \tilde{\nu}) \\ \sin(\psi_i - \tilde{\nu}) & -\cos(\psi_i - \tilde{\nu}) \end{bmatrix} \right) \\
&+ \begin{bmatrix} \cos(\mu_0 - \tilde{\mu}) & \sin(\mu_0 - \tilde{\mu}) \\ \sin(\mu_0 - \tilde{\mu}) & -\cos(\mu_0 - \tilde{\mu}) \end{bmatrix} A_0' \begin{bmatrix} \cos(\nu_0 - \tilde{\nu}) & \sin(\nu_0 - \tilde{\nu}) \\ \sin(\nu_0 - \tilde{\nu}) & -\cos(\nu_0 - \tilde{\nu}) \end{bmatrix}
\end{aligned}
$$

where $A_0' = \begin{bmatrix} a & -b \\ -c & d \end{bmatrix}$. Note that when $b = c = 0$, as in the case of the Rivest (1988), sine (Singh et al., 2002), and cosine (Mardia et al., 2007) models, then $A_0 = A_0'$.

So if we are dealing with a bivariate von Mises sine model with a sine model prior, in which case A_i is a matrix with λ_i in the lower right corner and 0s elsewhere, then:

$$\begin{aligned}\tilde{A} &= \sum_{i=0}^{n} \begin{bmatrix} \cos(\phi_i - \tilde{\mu}) & \sin(\phi_i - \tilde{\mu}) \\ \sin(\phi_i - \tilde{\mu}) & -\cos(\phi_i - \tilde{\mu}) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \lambda_i \end{bmatrix} \begin{bmatrix} \cos(\psi_i - \tilde{\nu}) & \sin(\psi_i - \tilde{\nu}) \\ \sin(\psi_i - \tilde{\nu}) & -\cos(\psi_i - \tilde{\nu}) \end{bmatrix} \\ &= \sum_{i=0}^{n} \lambda_i \begin{bmatrix} \sin(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & -\sin(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \\ -\cos(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & \cos(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \end{bmatrix}\end{aligned}$$

An alternative derivation for the full conditional distribution for the sine model was independently developed by Mardia (2009).

APPENDIX B

DERIVATION OF FULL CONDITIONAL DISTRIBUTION IN CHAPTER III

As we showed in Appendix A, the full conditional distribution for a set of observations with bivariate von Mises sine model distributions and a sine model prior is an eight parameter bivariate von Mises distribution with parameters:

$$\tilde{\mu} = \arctan\left(\sum_{i=0}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right) \qquad \tilde{\nu} = \arctan\left(\sum_{i=0}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right)$$
$$\tilde{\kappa}_1 = \left|\sum_{i=0}^{n} \kappa_{1i}[\cos(\phi_i), \sin(\phi_i)]\right| \qquad \tilde{\kappa}_2 = \left|\sum_{i=0}^{n} \kappa_{2i}[\cos(\psi_i), \sin(\psi_i)]\right|. \qquad \text{(A.1)}$$
$$\tilde{A} = \sum_{i=0}^{n} \lambda_i \begin{bmatrix} \sin(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & -\sin(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \\ -\cos(\phi_i - \tilde{\mu})\sin(\psi_i - \tilde{\nu}) & \cos(\phi_i - \tilde{\mu})\cos(\psi_i - \tilde{\nu}) \end{bmatrix}$$

where C is the appropriate constant of integration and the prior mean parameters (μ_0, ν_0) are treated as an additional observation (ϕ_0, ψ_0) from a bivariate von Mises sine model with parameters μ, ν, κ_{10}, κ_{20}, and λ_0.

Now consider a prior distribution of the form:

$$\pi(\mu, \nu) = \sum_{k=1}^{K} p_k C_k \exp\{\kappa_{10k}\cos(\mu_{0k} - \mu) + \kappa_{20k}\cos(\nu_{0k} - \nu) + \lambda_{0k}\sin(\mu_{0k} - \mu)\sin(\nu_{0k} - \nu)\},$$

where C_k is the constant of integration for a von Mises sine model with parameters κ_{10k}, κ_{20k}, and λ_{0k} given in equation (3.2), $p_k \geq 0$ for $k = 1, ...K$, and $\sum_{k=1}^{K} p_k = 1$. The full conditional distribution is proportional to this distribution times the likelihood, giving:

$$\begin{aligned}\pi(\mu,\nu|\boldsymbol{\phi},\boldsymbol{\psi}) \propto L(\mu,\nu|\boldsymbol{\phi},\boldsymbol{\psi}) \sum_{k=1}^{K} p_k C_k \exp\{&\kappa_{10k}\cos(\mu_{0k}-\mu)+\kappa_{20k}\cos(\nu_{0k}-\nu)\\ &+\lambda_{0k}\sin(\mu_{0k}-\mu)\sin(\nu_{0k}-\nu)\}\\ =\sum_{k=1}^{K} p_k L(\mu,\nu|\boldsymbol{\phi},\boldsymbol{\psi}) C_k \exp\{&\kappa_{10k}\cos(\mu_{0k}-\mu)+\kappa_{20k}\cos(\nu_{0k}-\nu)\\ &+\lambda_{0k}\sin(\mu_{0k}-\mu)\sin(\nu_{0k}-\nu)\},\end{aligned}$$

where $L(\mu,\nu|\boldsymbol{\phi},\boldsymbol{\psi})$ is the likelihood excluding the constant of integration.

Each term in the sum depends on the unknown parameters only through the product of the likelihood and a single von Mises sine distribution. This product is proportional to an eight parameter bivariate von Mises distribution with parameters given by (A.1). Call the resulting posterior parameters $\tilde{\mu}_i$, $\tilde{\nu}_i$, and so on. Then the full conditional distribution is proportional to:

$$\sum_{k=1}^{K} p_k C_k \exp\{\tilde{\kappa}_{1k}\cos(\mu-\tilde{\mu}_k)+\tilde{\kappa}_{2k}\cos(\nu-\tilde{\nu}_k)+[\cos(\mu-\tilde{\mu}),\sin(\mu-\tilde{\mu})]\tilde{A}_k[\cos(\mu-\tilde{\mu}),\sin(\nu-\tilde{\nu})]^T,$$

which integrates to:

$$\sum_{k=1}^{K} p_k C_k \tilde{C}_k^{-1},$$

where $\tilde{C}_k$ is the constant of integration for an eight parameter bivariate von Mises distribution with parameters $\tilde{\mu}_k$, $\tilde{\nu}_k$, $\tilde{\kappa}_{1k}$, $\tilde{\kappa}_{2k}$, and $\tilde{\lambda}_k$. Therefore, the full conditional distribution takes the form:

$$\pi(\mu,\nu|\boldsymbol{\phi},\boldsymbol{\psi}) = \sum_{k=1}^{K} p_k^* f(\mu,\nu|\tilde{\mu}_k,\tilde{\nu}_k,\tilde{\kappa}_{1k},\tilde{\kappa}_{2k},\tilde{A}_k),$$

where f is an eight parameter bivariate von Mises distribution and $p_k^* = (p_k C_k \tilde{C}_k^{-1})/(\sum_{j=1}^{K} p_j C_j \tilde{C}_j^{-1})$. Note that $p_k^* \geq 0$ for $k = 1, ..., K$, and $\sum_{k=1}^{K} p_k^* = 1$.

Unfortunately computational formulas for the constant of integration of a bivariate von Mises distribution do not exist in the general case. Therefore we do not sample directly from this full conditional distribution, but rather use an independence sampler which replaces each full conditional eight parameter distribution with a five parameter sine model, and uses the corresponding constant of integration from (3.2). For this sine model based proposal distribution we keep the true full conditional mean and precision parameters, and take $\tilde{\lambda} = \left(\sum_{i=0}^{n} \lambda_i \cos(\phi_i - \psi_i)\right) \left\{\cos(\tilde{\mu} - \tilde{\nu})\right\}^{-1}$. This method is a direct extension of the single sine model prior case presented in Chapter II.

VITA

Name: Kristin Patricia Lennox

Address: Department of Statistics
Texas A&M University
3143 TAMU
College Station, TX 77843-3143

E-mail: lennox@stat.tamu.edu

Education: B.Sci. Applied Mathematics, Texas A&M University, 2005
M.Sci. Statistics, Texas A&M University, 2007
Ph.D. Statistics, Texas A&M University, 2010