Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics

Kristin P. LENNOX, David B. DAHL, Marina VANNUCCI, and Jerry W. TSAI

Interest in predicting protein backbone conformational angles has prompted the development of modeling and inference procedures for bivariate angular distributions. We present a Bayesian approach to density estimation for bivariate angular data that uses a Dirichlet process mixture model and a bivariate von Mises distribution. We derive the necessary full conditional distributions to fit the model, as well as the details for sampling from the posterior predictive distribution. We show how our density estimation method makes it possible to improve current approaches for protein structure prediction by comparing the performance of the so-called "whole" and "half" position distributions. Current methods in the field are based on whole position distributions, as density estimation for the half positions requires techniques, such as ours, that can provide good estimates for small datasets. With our method we are able to demonstrate that half position data provides a better approximation for the distribution of conformational angles at a given sequence position, therefore providing increased efficiency and accuracy in structure prediction.

KEY WORDS: Angular data; Density estimation; Dirichlet process mixture model; Torsion angles; von Mises distribution.

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, Ozkan, Weikl, Chodera, and Voelz 2007).

Based on the fundamental work of Ramachandran, Ramakrishnan, and Sasisekharan (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) (Fig. 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 toward predicting conformational angles for novel proteins (Ho, Thomas, and Brasseur 2003; Xue, Dor, Faraggi, and Zhou 2008).

Datasets from the Protein Structure Databank (PDB) (Berman et al. 2000) can consist of over ten thousand angle pairs, which provide ample data for even relatively unsophisticated density 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, Zhou, and Ohlson 2002; Lovell et al. 2003; Rother, Sapiro, and Pande 2008), but they behave poorly for these subdivided datasets. This is unfortunate, because these subsets provide structure prediction that is more accurate, as it 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, Zelinka, Fonson, Henderson, and Otwinowski (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 that 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 expectation-maximization (EM) algorithm. Dahl, Bohannan, Mo, Vannucci, and Tsai (2008) used 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 nonangular distribution, also provides estimates that properly account for the wrapped nature of angular data. In addition, the

Kristin P. Lennox is a doctoral candidate, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *lennox@stat.tamu.edu*). David B. Dahl is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *dahl@stat.tamu.edu*). Marina Vannucci is Professor, Department of Statistics, Rice University, Houston, TX 77251 (E-mail: *marina@rice.edu*). Jerry W. Tsai is Associate Professor, Department of Chemistry, University of the Pacific, 3601 Pacific Avenue, Stockton, CA 95211 (E-mail: *jtsai@pacific.edu*). This work was supported by an NIH/NIGMS grant R01GM81631. Dr. Vannucci is also supported by NIH/NHGRI grant R01HG003319 and by NSF/DMS grant number DMS-0605001. The authors thank the associate editor and referees for helpful suggestions, as well as J. Bradley Holmes, Jerod Parsons, and Kun Wu for help with datasets, alignments, and the torsion angle calculations.

© 2009 American Statistical Association
Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Applications and Case Studies
DOI 10.1198/jasa.2009.0024

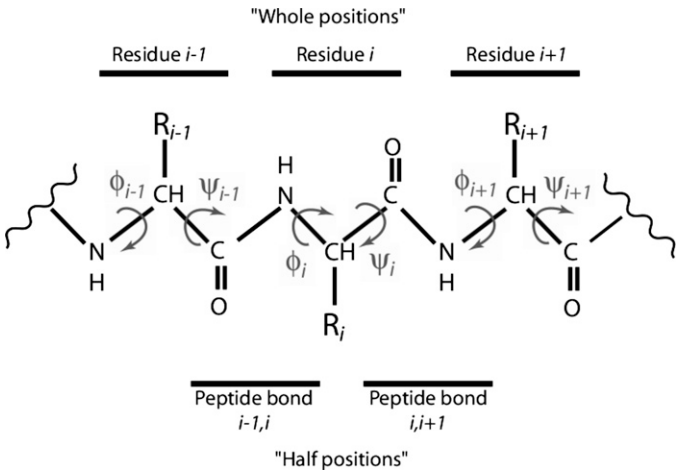

Figure 1. Diagram of protein backbone, including ϕ and ψ angles, whole positions, and half positions. At the ith residue the ϕ angle describes the torsion around the bond N_i–$C_{\alpha i}$, measuring the angle between the C_{i-1}–N_i and the $C_{\alpha i}$–C_i bonds, whereas the ψ angle describes the 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_α 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.

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). Because of the complexity of this distribution, methods for sampling from the posterior distribution are not obvious. Therefore, we provide a Markov chain Monte Carlo (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 a protein 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, because these angle pairs are associated with two adjacent residue types, as opposed to a single residue for whole positions. Because 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. Because of their specificity, datasets for half positions are often relatively small, a situation that our proposed density estimation technique handles well.

Section 2 of this article contains a review of past work in angular data analysis, including recent work in mixture modeling. In Section 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 4 we also 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 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 6.

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.

2.1 Univariate Angular Data

A common option for describing univariate circular data are the von Mises distribution (e.g., see 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, whereas 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, Leite, and Milan (2000) presented an empirical Bayes approach to inference.

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:

$$f(\phi,\psi|\mu,\nu,\kappa_1,\kappa_2,\lambda) = C\ \exp\{\kappa_1\cos(\phi-\mu) + \kappa_2\cos(\psi-\nu) + \lambda\sin(\phi-\mu)\sin(\psi-\nu)\} \quad (1)$$

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

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

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 $\Sigma_{11}^{-1} = \kappa_1, \Sigma_{22}^{-1} = \kappa_2$, and $\Sigma_{12}^{-1} = \Sigma_{21}^{-1} = -\lambda$. 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.

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

$$(\phi_i,\psi_i)|\mu_i,\nu_i,\Omega_i \sim p((\phi_i,\psi_i)|\mu_i,\nu_i,\Omega_i) \quad (3)$$

$$(\mu_i,\nu_i,\Omega_i)|G \sim G \quad (4)$$

$$G \sim DP(\tau_0 H_1 H_2), \quad (5)$$

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 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 resulting model 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 took 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, the Dahl et al. (2008) model 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 was first discussed by Escobar and West (1995). A nonparametric density estimate of the (ϕ, ψ) space from data $(\phi, \psi) = ((\phi_1, \psi_1), \ldots, (\phi_n, \psi_n))$ is the posterior predictive distribution of a new angle pair (ϕ_{n+1}, ψ_{n+1}), namely:

$$\begin{aligned} &p((\phi_{n+1},\psi_{n+1})|(\boldsymbol{\phi},\boldsymbol{\psi})) \\ &= \int p((\phi_{n+1},\psi_{n+1}),(\mu_{n+1},\nu_{n+1},\Omega_{n+1})|(\boldsymbol{\phi},\boldsymbol{\psi}))d(\mu_{m+1},\nu_{n+1},\Omega_{n+1}) \\ &= \int p((\phi_{n+1},\psi_{n+1})|(\mu_{n+1},\nu_{n+1},\Omega_{n+1})) \\ &\quad \times p((\mu_{n+1},\nu_{n+1},\Omega_{n+1})|(\boldsymbol{\phi},\boldsymbol{\psi}))d(\mu_{n+1},\nu_{n+1},\Omega_{n+1}). \end{aligned} \quad (6)$$

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

4. MODEL ESTIMATION

The integral of the posterior predictive density in (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

$$p((\phi_{n+1},\psi_{n+1})|(\boldsymbol{\phi},\boldsymbol{\psi})) \approx \frac{1}{B}\sum_{b=1}^{B} p((\phi_{n+1},\psi_{n+1})|(\mu_{n+1}^b,\nu_{n+1}^b,\Omega_{n+1}^b)). \quad (7)$$

Although Equation (7) can be evaluated for any value of (ϕ_{n+1}, ψ_{n+1}), for our purposes we obtain density estimates by evaluating (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 objects to clusters. We are at liberty to choose any valid updating scheme for the mean and precision parameters. Because 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) sampling methods for each full conditional distribution.

4.1 Full Conditional Distributions of Mean and Precision Parameters

Note that the sine model can be written either in terms of angles ϕ_i and ψ_i, or in terms of coordinates on the unit circle $x_i = (\cos(\phi_i), \sin(\phi_i))^T$, and $y_i = (\cos(\psi_i), \sin(\psi_i))^T$. Let $\tau_\mu = (\cos(\mu), \sin(\mu))^T$, $\tau_\nu = (\cos(\nu), \sin(\nu))^T$, and for a vector $\delta = (a, b)^T$, let $\delta^*: = (-b, a)^T$. The sine model density from (1) can now be rewritten as

$$f(x_i, y_i|\tau_\mu, \tau_\nu, \kappa_1, \kappa_2, \lambda) = C\exp(\kappa_1\tau_\mu^T x_i + \kappa_2\tau_\nu^T y_i + \lambda\tau_\mu^{*T} x_i \tau_\nu^{*T} y_i).$$

This formulation is used in the derivation of the full conditional distribution of the mean parameters given in the Appendix.

The conditionally conjugate prior for the mean parameters is a von Mises sine model with center $x_0 = (\cos(\mu_0), \sin(\mu_0))^T$, $y_0 = (\cos(\nu_0), \sin(\nu_0))^T$, and precision parameters κ_{10}, κ_{20}, and λ_0. Consider a set of observations (x_i, y_i), $i = 1, \ldots, n$, each with known precision parameters κ_{1i}, κ_{2i}, and λ_i. The full conditional distribution is a von Mises sine model with parameters:

$$\tilde{\mu} = \arctan\left(\sum_{i=0}^{n} \kappa_{1i} x_i\right) \quad \tilde{\nu} = \arctan\left(\sum_{i=0}^{n} \kappa_{2i} y_i\right)$$

$$\tilde{\kappa}_1 = \left|\sum_{i=0}^{n} \kappa_{1i} x_i\right| \quad \tilde{\kappa}_2 = \left|\sum_{i=0}^{n} \kappa_{2i} y_i\right|$$

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

The mean parameters of the full conditional distribution are the directions of the sums of the observation vectors, whereas 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).

The conditionally conjugate 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 a noninformative alternative prior to the conditionally conjugate is a uniform distribution on $(-\pi, \pi] \times (-\pi, \pi]$, which is the limit of the sine model prior when $\lambda_0 = 0$ and $\kappa_{10}, \kappa_{20} \to 0$.

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

$$\pi(\kappa_1, \kappa_2, \lambda) \propto \left\{4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_m(\kappa_2)\right\}^{-c} \times \exp(R_{\phi 0}\kappa_1 + R_{\psi 0}\kappa_2 + R_{\phi\psi 0}\lambda). \quad (8)$$

Here the prior assumes the role of c observations from the von Mises distribution, 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 all fall between $-c$ and c. Notice that the conditionally conjugate prior, and corresponding posterior distribution, are difficult to sample from because of the infinite sums of Bessel functions. Notice also that these distributions do not guarantee precision parameters that will give unimodal sine model distributions.

4.2 Markov Chain Monte Carlo Sampler

The posterior distribution of our model's parameters in Section 3 can be sampled using MCMC via the Auxiliary Gibbs sampler of Neal (2000). This method requires the ability to directly sample from the centering distribution. Because it is difficult to sample from the conjugate prior for the precision parameters described in (8), we instead use the Wishart distribution for $H_2(\Omega)$ in (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 (8).

Auxiliary Gibbs sampling requires a valid updating scheme for the model parameters. Direct sampling from the full conditional distribution of the means is fairly straightforward. As described previously in Section 4.1, the full conditional distribution of the means is given by a von Mises sine model. A simple method to sample from this distribution 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 $(\tilde{\mu}, \tilde{\nu})$, whereas the values in the bimodal case depend on the sign of $\tilde{\lambda}$ and are given in the Appendix of Mardia et al. (2007).

An 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 because of 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. Furthermore, 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).

Computational Procedure

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 full conditional distribution in Section 4.1.
   (c) Given the clustering configuration and mean values, update the precision matrix Ω for each cluster using the Wishart independence sampler described in Section 4.2.

5. TEMPLATE BASED MODELING OF PROTEIN STRUCTURE

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 because consideration of the periodicity is essential for certain amino acids, such as glycine. Figure 2 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 20 amino acid datasets, once using normal centering 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 the normal model estimate. The logs of the Bayes factors ranged from 216 to 5,040 in absolute value, allowing us to draw clear conclusions as to the superior model in each case. For 14 of the 20 amino acids, the Bayes factor indicated that the von Mises model was superior. When we restricted our test set to only amino acids with more that 2.5% of the observations within $\pi/18$ radians (10 degrees) of the border of the Ramachandran plot, 11 of the 12 datasets had a better fit with the von Mises model. Although 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 because 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

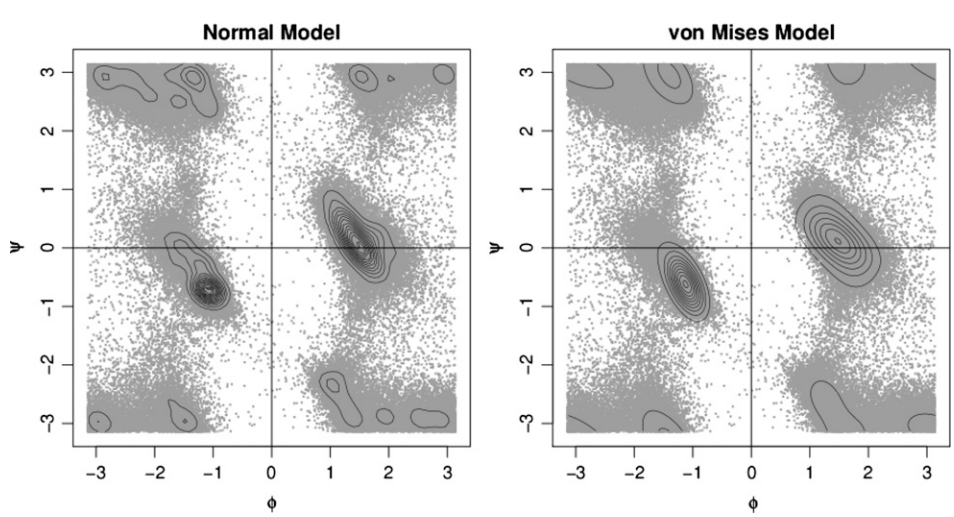

Figure 2. Ramachandran plots for the 121,497 angle pairs that make up the PDB dataset for the residue glycine, along with density estimates based on both the normal and von Mises distributions. The normal model is from the work of Dahl et al. (2008), whereas the von Mises estimate is based on our model in Section 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.

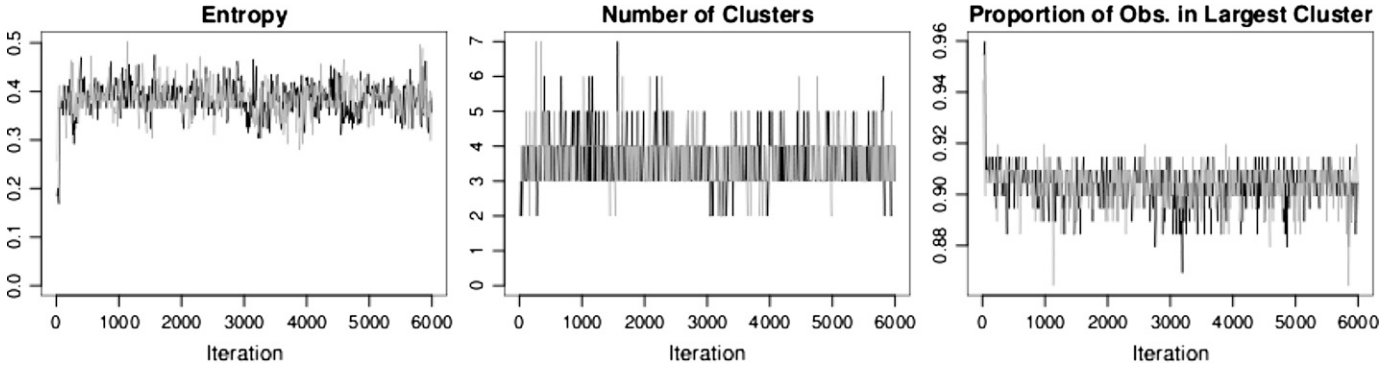

Figure 3. Convergence diagnostics for threonine-arginine half position dataset.

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 threefold families: globins, immunoglobulins, and triose phosphate isomerase (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.

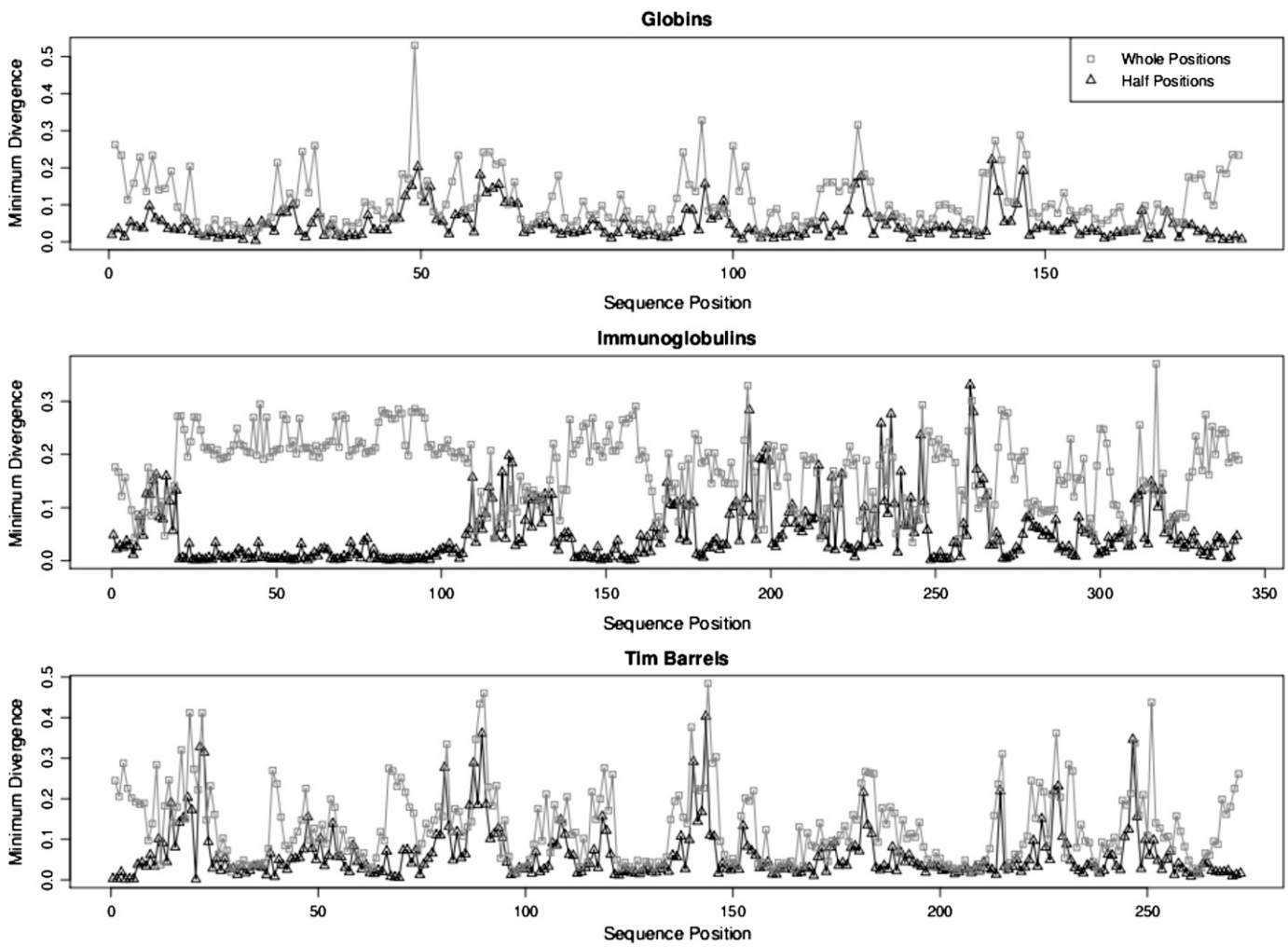

Figure 4. A comparison of minimum divergence scores for whole versus half positions.

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 Fig. 1). Previously, this peptide centered view of torsion angles has only been applied to short peptides (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,'' because 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 20 amino acid residues is present, and also the type of secondary structure

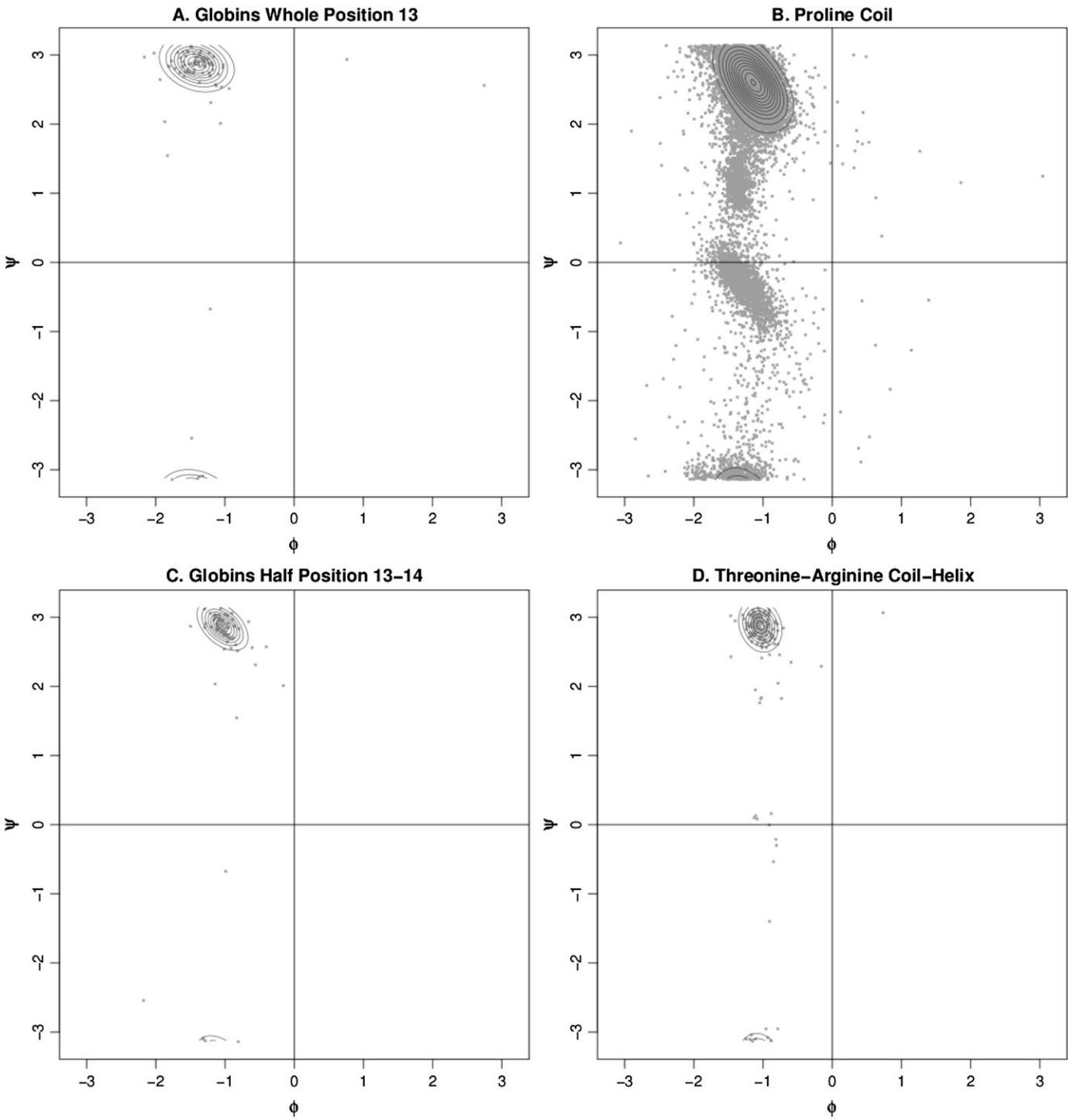

Figure 5. Ramachandran plots around globins position 13. (*a*) Density estimate and data for whole position 13. (*b*) Density estimate and data for proline coil whole positions which, at a divergence of 0.204, provide the best PDB match for globins whole position 13. (*c*) Density estimate and data for half position between residues 13 and 14. (d) Density estimate and data for threonine coil to arginine helix half positions which, at a divergence of 0.028, provide the best PDB match for globins half position 13–14.

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. The 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 20 residues and 4 secondary structure classes provide 80 possible classifications for whole position data.

Because a half position involves two residues, there are 400 categories when considering only amino acid pairs and 6,400 when the 4 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, because 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.

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 globins 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 (7). For datasets with over 2,000 observations, one run of the MCMC sampler used the full dataset, whereas the other used a subsample of 2,000 observations. We found that subsampling had little impact on the density estimates.

Our von Mises model from Section 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 $\upsilon = 2$ 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 $\upsilon/2B^{-1} = B^{-1}$). 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 3 shows trace plots for the threonine-arginine half position, which provided the best match from the PDB data for the globins 13–14 half position. Notice the rapid convergence for all of our univariate convergence criteria, which was typical for our MCMC scheme.

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 $D_{KL}(P, Q) = \sum_i P(i)\log(P(i)/Q(i))$. Both P and Q are density estimates from our proposed procedure.

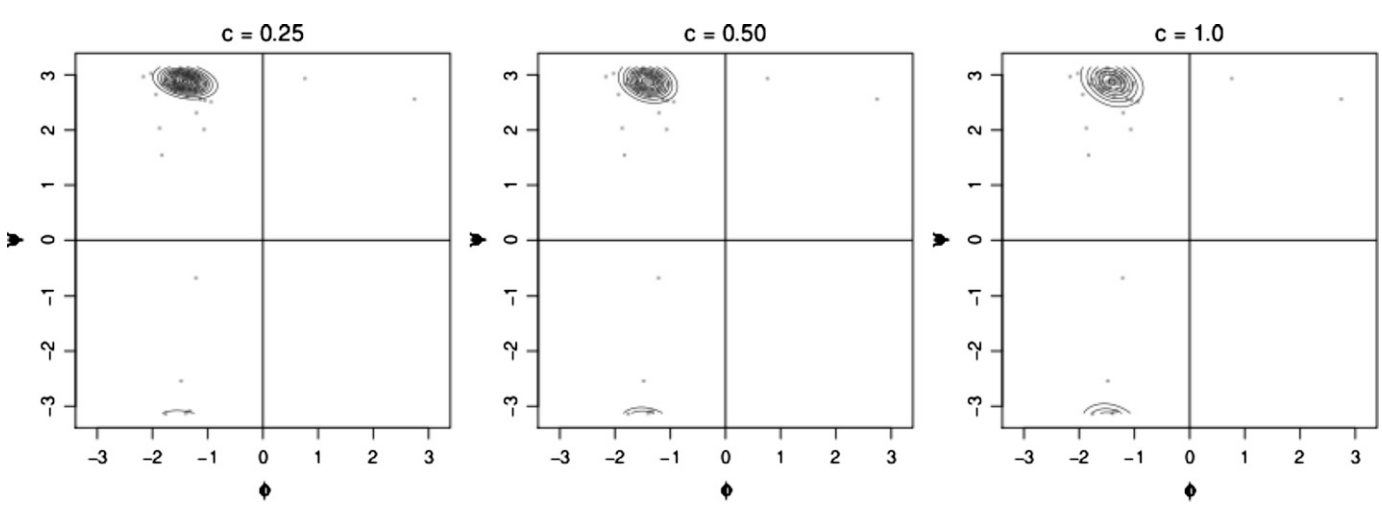

Figure 6. Density estimates for globins positon 13 with different scale matrices for the Wishart prior distribution.

The positional density estimates were compared with all of the estimates from the PDB using this divergence score. Whole position densities from each of the three fold families were compared with the whole position category densities from the PDB, and half positions from the fold families were compared with the half position category densities from the PDB. The best matches, those with the lowest divergence values, are plotted against position in Figure 4. It is evident that the half position comparisons produce lower divergence scores. The mean minimum divergence for whole positions is 0.145, whereas the corresponding half position value is 0.054. 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 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 with a mean divergence of 0.066. Even then, half positions provide a better match for helices with mean divergence 0.031. The worst matching cases are in areas with noncanonical 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 5, which shows the globins whole position 13 with the closest matching PDB density compared with the half position 13–14 with its matching half position density from the PDB. It can be readily seen from these 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 because of the increased specificity of the half position densities from the PDB, not to mention the increased number of categories available for comparison. For the globins in Figure 5, the whole position density certainly shows a larger cross section of density than the half position. These results suggest that the use of half position data as a substitute for whole position data provides better results.

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

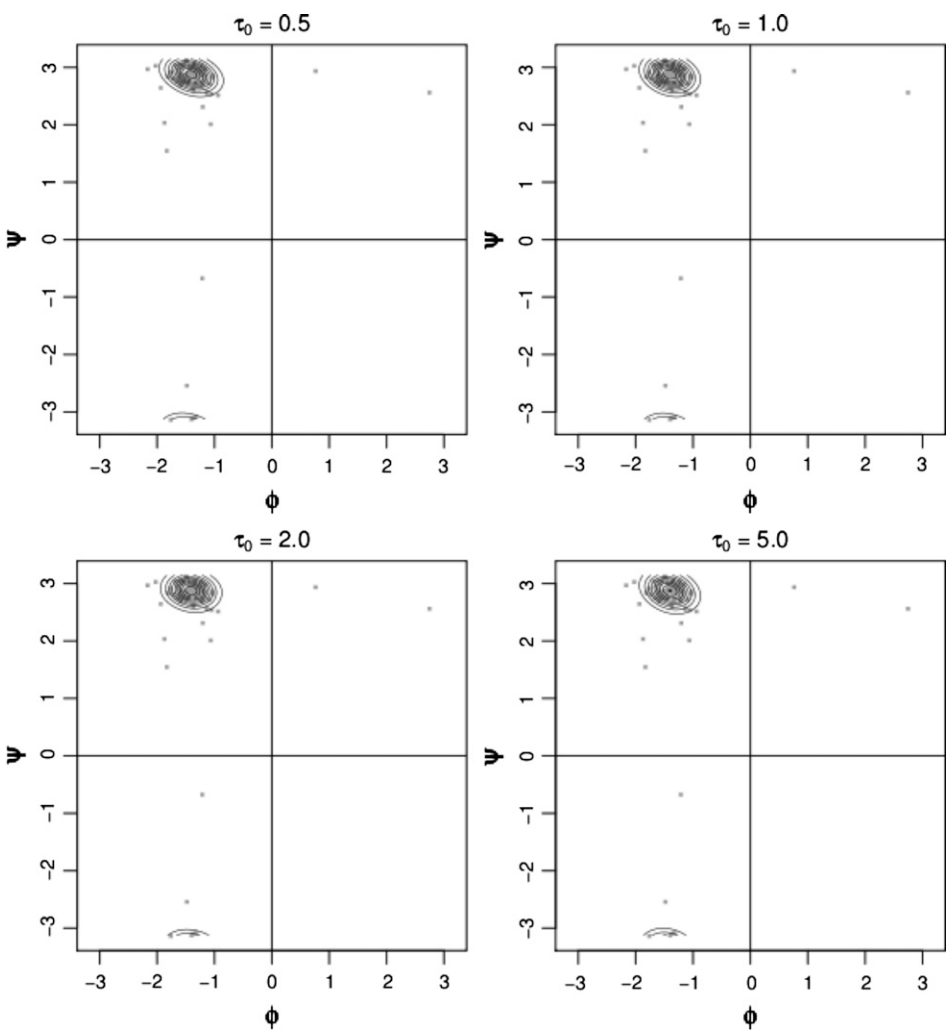

Figure 7. Density estimates for globins position 13 for assorted values of the mass parameter.

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 MCMC 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 shown here, most protein folds have very limited number 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 6 shows the resulting density estimates for globins position 13. The changes between the density estimates are not dramatic, and the effect is comparable to that of varying the bandwidth in kernel density estimation methods. Other positions showed similar behavior.

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 13 is given in Figure 7. 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, although the trace plots of entropy for the two chains might suggest convergence problems, the density estimates generated by the separate chains were very similar. Therefore, we do not consider this to be a major issue. On the other hand, if the mass parameter is very small severe convergence problems can occur. As always, convergence diagnostics should be employed.

APPENDIX: DERIVATION OF RELEVANT DISTRIBUTIONS

Here we provide the derivations of the results from Section 4.1 regarding the conditionally conjugate prior and posterior distribution for the case when we have data from a bivariate von Mises sine model with unknown mean parameters and known precision parameters.

Recall the parameterization of the von Mises in terms of unit vectors from Section 4.1, where $x_i = (\cos(\phi_i), \sin(\phi_i))^T$, $y_i = (\cos(\psi_i), \sin(\psi_i))^T$, $\tau_\mu = (\cos(\mu), \sin(\mu))^T$, $\tau_\nu = (\cos(\nu), \sin(\nu))^T$, and for a vector $\delta = (a, b)^T$, $\delta^* := (-b, a)^T$. For observations (x_i, y_i), $i = 1, \ldots, n$ from a bivariate von Mises distribution with known κ_{1i}, κ_{2i}, and λ_i, the likelihood for τ_μ and τ_ν is

$$\begin{aligned} L(\tau_\mu, \tau_\nu | \mathbf{x}, \mathbf{y}) &\propto \prod_{i=1}^n \exp(\kappa_{1i}\tau_\mu^T x_i + \kappa_{2i}\tau_\nu^T y_i + \lambda_i \tau_\mu^{*T} x_i \tau_\nu^{*T} y_i) \\ &= \exp\left\{ \tau_\mu^T \sum_{i=1}^n \kappa_{1i} x_i + \tau_\nu^T \sum_{i=1}^n \kappa_{2i} y_i \right. \\ &\quad \left. + \sum_{i=1}^n (\lambda_i \tau_\mu^{*T} x_i \tau_\nu^{*T} y_i) \right\}. \end{aligned}$$

Notice that $\tau_\mu^{*T} x_i \tau_\nu^{*T} y_i$ is equal to $\tau_\mu^{*T} \tau_\nu^* x_i^T y_i$, and that for all vectors a, b, we have $a^{*T} b^* = a^T b$, so

$$\begin{aligned} L(\tau_\mu, \tau_\nu | \mathbf{x}, \mathbf{y}) &\propto \exp\left\{ \left(\sum_{i=1}^n \kappa_{1i} x_i \right)^T \tau_\mu + \left(\sum_{i=1}^n \kappa_{2i} y_i \right)^T \tau_\nu \right. \\ &\quad \left. + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) \tau_\mu^T \tau_\nu \right\} \\ &= \exp\left\{ \left| \sum_{i=1}^n \kappa_{1i} x_i \right| \frac{\left(\sum_{i=1}^n \kappa_{1i} x_i\right)^T}{\left|\sum_{i=1}^n \kappa_{1i} x_i\right|} \tau_\mu \right. \\ &\quad + \left| \sum_{i=1}^n \kappa_{2i} y_i \right| \frac{\left(\sum_{i=1}^n \kappa_{2i} y_i\right)^T}{\left|\sum_{i=1}^n \kappa_{2i} y_i\right|} \tau_\nu \\ &\quad \left. + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) \tau_\mu^T \tau_\nu \right\}. \end{aligned}$$

Notice that the first two terms are consistent with a von Mises likelihood with

$$\tilde{\tau}_\mu = \frac{\left(\sum_{i=1}^n \kappa_{1i} x_i\right)}{\left|\sum_{i=1}^n \kappa_{1i} x_i\right|} \quad \tilde{\tau}_\nu = \frac{\left(\sum_{i=1}^n \kappa_{2i} y_i\right)}{\left|\sum_{i=1}^n \kappa_{2i} y_i\right|}$$

$$\tilde{\kappa}_1 = \left| \sum_{i=1}^n \kappa_{1i} x_i \right| \quad \tilde{\kappa}_2 = \left| \sum_{i=1}^n \kappa_{2i} y_i \right|.$$

Using this notation, we have

$$\begin{aligned} L(\tau_\mu, \tau_\nu | \mathbf{x}, \mathbf{y}) &\propto \exp\left\{ \tilde{\kappa}_1 \tilde{\tau}_\mu^T \tau_\mu + \tilde{\kappa}_2 \tilde{\tau}_\nu^T \tau_\nu + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) \tau_\mu^T \tau_\nu \right\} \\ &= \exp\left\{ \tilde{\kappa}_1 \tilde{\tau}_\mu^T \tau_\mu + \tilde{\kappa}_2 \tilde{\tau}_\nu^T \tau_\nu + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) \tau_\mu^T \tau_\nu \tilde{\tau}_\mu^T \tilde{\tau}_\nu (\tilde{\tau}_\mu^T \tilde{\tau}_\nu)^{-1} \right\} \\ &= \exp\left\{ \tilde{\kappa}_1 \tilde{\tau}_\mu^T \tau_\mu + \tilde{\kappa}_2 \tilde{\tau}_\nu^T \tau_\nu + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) (\tilde{\tau}_\mu^T \tilde{\tau}_\nu)^{-1} \tau_\mu^T \tau_\nu \tilde{\tau}_\mu^{*T} \tilde{\tau}_\nu^* \right\} \\ &= \exp\left\{ \tilde{\kappa}_1 \tilde{\tau}_\mu^{\mathrm{T}} \tau_\mu + \tilde{\kappa}_2 \tilde{\tau}_\nu^T \tau_\nu + \left(\sum_{i=1}^n \lambda_i x_i^T y_i \right) (\tilde{\tau}_\mu^T \tilde{\tau}_\nu)^{-1} \tilde{\tau}_\mu^{*T} \tau_\mu \tilde{\tau}_\nu^{*T} \tau_\nu \right\}. \end{aligned}$$

So, in addition to the parameters defined previously, the likelihood is proportional to a von Mises sine model with

$$\tilde{\lambda} = \sum_{i=1}^{n} \left(\lambda_i x_i^T y_i\right)\left(\tilde{\tau}_\mu{}^T \tilde{\tau}_\nu\right)^{-1}.$$

The conjugate prior can be treated as an additional observation in the likelihood, and its interpretation follows directly.

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.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000), "The Protein Data Bank," *Nucleic Acids Research,* 28, 235–242.

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.

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.

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.

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," *Scandinavian Journal of Statistics,* 28, 355–375.

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.

Hovmoller, S., Zhou, T., and Ohlson, T. (2002), "Conformations of Amino Acids in Proteins," *Acta Crystallographica. Section D, Biological Crystallography,* 58, 768–776.

Kabsch, W., and Sander, C. (1983), "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features," *Biopolymers,* 22, 2577–2637.

Lee, S. Y., and Skolnick, J. (2008), "Benchmarking of TASSER 2.0: An Improved Protein Structure Prediction Algorithm with More Accurate Predicted Contact Restraints," *Biophysical Journal,* 95, 1956–1964.

Lovell, S. C., Davis, I. W., Arendall, W. B.r., 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.

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., and El-Atoum, S. A. M. (1976), "Bayesian Inference for the Von Mises-Fisher Distribution," *Biometrika,* 63, 203–205.

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.

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.

Pertsemlidis, A., Zelinka, J., Fondon, J. W.r., Henderson, R. K., and Otwinowski, Z. (2005), "Bayesian Statistical Studies of the Ramachandran Distribution." in *Statistical Applications in Genetics and Molecular Biology,* (Vol. 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. (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.

Rother, D., Sapiro, G., and Pande, V. (2008), "Statistical Characterization of Protein Ensembles," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 45, 42–55.

Singh, H., Hnizdo, V., and Demchuk, E. (2002), "Probabilistic Model for Two Dependent Circular Variables," *Biometrika,* 89, 719–723.

Xue, B., Dor, O., Faraggi, E., and Zhou, Y. (2008), "Real-Value Prediction of Backbone Torsion Angles," *Proteins,* 72, 427–433.

Correction to "Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics" [JASA 104, p. 586-596]

The full conditional distribution for the mean parameters (μ, ν) of a bivariate von Mises sine model is incorrectly specified in Section 4.1 and the Appendix. Rather than a five parameter bivariate von Mises sine model with correlation parameter $\tilde{\lambda}$, the full conditional distribution is an eight parameter bivariate von Mises distribution. Using the representation from Mardia et al. (2007), the correct full conditional distribution can be expressed as:

$$f(\mu, \nu | \boldsymbol{\phi}, \boldsymbol{\psi}) \propto \exp\{\tilde{\kappa}_1 \cos(\mu - \tilde{\mu}) + \tilde{\kappa}_2 \cos(\nu - \tilde{\nu}) + [\cos(\mu - \tilde{\mu}), \sin(\mu - \tilde{\mu})]\tilde{A}[\cos(\nu - \tilde{\nu}), \sin(\nu - \tilde{\nu})]^T\}$$

where $\tilde{\mu}$, $\tilde{\nu}$, $\tilde{\kappa}_1$, and $\tilde{\kappa}_2$ are specified correctly in the paper, and

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

We wish to acknowledge Professor Kanti Mardia, who also independently discovered the error and the correct distribution (Mardia 2009).

In making this correction, we found the five parameter von Mises sine model given in our paper to be a good proposal distribution for a Metropolis-Hastings update of (μ, ν). For our corrected analysis of the whole and half position data the mean and median for the acceptance rate were around 60%, and our results were not substantially affected. For example, Figure 4 showed almost identical minimum divergence values, and thus the same improvement for the half position estimates with respect to the whole position estimates. More details on the corrected derivation and results are provided in a supplemental file.

Mardia, K. V. (2009), "Statistical Complexity in Protein Bioinformatics," *LASR 2009 Proceedings*, Editors A. Gustano, K. V. Mardia and C. J. Fallaize, July 7 - July 9, 2009, Leeds, Leeds University Press, 9-20.

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.

Supplemental File for Correction to "Density Estimation for Protein Conformation Angles Using a Bivariate von Mises Distribution and Bayesian Nonparametrics" JASA 104, p. 586–596

Kristin P. Lennox David B. Dahl Marina Vannucci Jerry W. Tsai

1 Corrected Derivation of Full Conditional Distribution of (μ, ν)

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:

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

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:

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

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:

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

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:

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

Mardia (2009) independently derived the full conditional distribution of the bivariate von Mises sine model and announced the general case.

2 Corrected Results

Originally in Section 5.1, the von Mises distribution outperformed the normal distribution for 14/20 amino acids (11/12 with high edge proportion). In the repeated analysis it outperformed the normal for 19/20 amino acids (11/12 with high edge proportion).

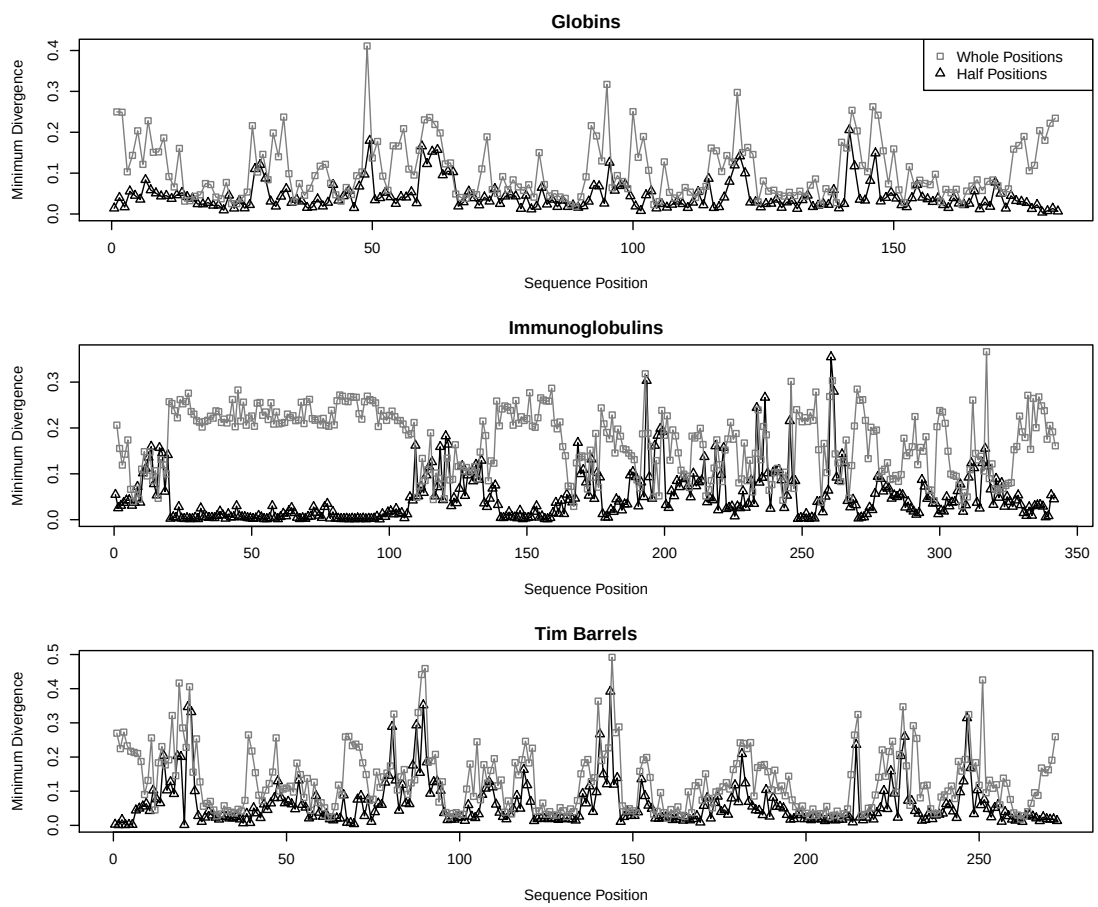

Figure 1: Corrected divergence plot.

The results of the corrected analysis for the whole and half position datasets were very similar to the original results. For example, the corrected divergence plot (originally Figure 4) is shown in Figure 1. Notice that the minimum divergence values are very similar to those presented in the paper, and the superiority of the half positions relative to the whole positions is unchanged.

References

Mardia, K. V. (2009), "Statistical Complexity in Protein Bioinformatics," *LASR 2009 Proceedings*, Editors A. Gustano, K. V. Mardia and C. J. Fallaize, July 7 - July 9, 2009, Leeds, Leeds University Press, 9-20.

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.

Rivest, L. P. (1988), "A Distribution for Dependent Unit Vectors," *Communications in Statistics: Theory and Methods*, 17, 461–483.

Singh, H., Hnizdo, V., and Demchuk, E. (2002), "Probabilistic Model for Two Dependent Circular Variables," *Biometrika*, 89, 719–723.