Protein Bioinformatics and Mixtures of Bivariate von Mises Distributions for Angular Data

Kanti V. Mardia & Charles C. Taylor
University of Leeds, Leeds, LS2 9JT

Ganesh K. Subramaniam
AT&T Labs Research

Abstract

A fundamental problem in bioinformatics is to characterize the secondary structure of a protein, which has traditionally been carried out by examining a scatter plot (Ramachandran plot) of the conformational angles. We examine two natural bivariate von Mises distributions — referred to as Sine and Cosine models — which have five parameters and, for concentrated data, tend to a bivariate normal distribution. These are analyzed and their main properties derived. Conditions on the parameters are established which result in bimodal behaviour for the joint density and the marginal distribution, and we note an interesting situation in which the joint density is bimodal, but the marginal distributions are unimodal. We carry out comparisons of the two models, and it is seen that the Cosine model may be preferred. Mixture distributions of the Cosine model are fitted to two representative protein datasets using the EM algorithm which results in an objective partition of the scatter plot into a number of components. Our results are consistent with emprirical observations; new insights are discussed.

Keywords: Bivariate angular data, Bivariate circular mixture, Directional statistics, Distribution on Torus, Myoglobin, Protein conformational angles, Ramachandran plots.

1 Introduction

Protein shapes are often described using the (3-d) co-ordinates of the backbone chain, which consists of an ordered sequence of nitrogen and carbon atoms known as peptide units. These co-ordinates can be used to obtain a sequence of conformational angles, known as *torsional* angles, which form a natural pairing (ϕ, ψ) with each angle in $(-\pi, \pi]$. These angles can be plotted in a scatter plot — now known as a Ramachandran plot (Ramachandran *et al.*, 1963) — and these have been used in the understanding of protein secondary structure, which describes a protein in terms of α helices, β-sheets and loop motifs etc.. Branden & Tooze (1998, p. 9) describe in more detail the uses of these very familiar plots, including an indication of the approximate segmentation into various motifs which was initially empirically observed by Ramachandran *et al.* (1963).

Recently, there have been attempts to parameterize the joint distribution of (ϕ, ψ) (Pertsemlidis *et al.*, 2005) and this has been done using Fourier basis functions to represent the bivariate distribution of (ϕ, ψ) on the torus. This parameterization requires about a hundred parameters and does not seem to allow an easy interpretation from a biological (or statistical) perspective.

One of our motivations for this paper is to describe the probability distribution on the torus, but our approach is to use a different parameterization which is based on a mixture of bivariate von Mises distributions. The von Mises distribution on the circle is well known; it has two parameters (analogous to the normal distribution), the mean (μ) and concentration parameter (κ) which is anti-variance. The density is given by (see, for example, Mardia & Jupp, 1999, p. 36)

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

where $I_0(\cdot)$ denotes the modified Bessel function of the first kind and order 0, and the parameter μ is the mean direction, and κ is the concentration parameter. This distribution can be approximated by the normal density for small “variance” (large κ). In Section 2, we introduce two bivariate von Mises distributions which, like the bivariate normal distribution, have 5 parameters: two means, two concentrations (anti-variance), and a parameter controlling “correlation”. In Section 3 we describe some key properties, and make some comparisons. In Section 4 we return to our motivating example of fitting an appropriate distribution to (ϕ, ψ) angles obtained from a protein backbone. The model is extended to mixtures of distributions which is fitted using an EM algorithm and some results for two proteins are described, A final discussion section describes further possible extensions using MCMC.

2 Bivariate von Mises Distributions

For the study of conformational angles (ϕ, ψ), we need a bivariate angular distribution, which extends the univariate von Mises distribution and has analogy with the bivariate normal distribution. Namely, two mean parameters, two parameters for variance (anti-variance) and a parameter which determines the correlation. Mardia (1975) introduced a bivariate circular distribution for (Φ, Ψ), with the probability density function (pdf) proportional to

$$\exp\left[\kappa_1\cos(\phi-\mu)+\kappa_2\cos(\psi-\nu)+\{\cos(\phi-\mu),\sin(\phi-\mu)\}A\{cos(\psi-\nu),\sin(\psi-\nu)\}^T\right] \tag{1}$$

where A is a 2×2 matrix. This has more parameters ($4+2+2=8$) than we would often require, there are difficulties of interpretation, and, at least for large concentrations, there is redundancy. A sub-class has been introduced by Rivest (1988) which has pdf proportional to

$$\exp\left\{\kappa_1\cos(\phi-\mu)+\kappa_2\cos(\psi-\nu)+\alpha\cos(\phi-\mu)\cos(\psi-\nu)+\beta\sin(\phi-\mu)\sin(\psi-\nu)\right\}. \tag{2}$$

However, both (1) and (2) are overparameterized in the sense that there should only be 5 parameters by analogy with the bivariate normal. To overcome this defect, Singh *et al.* (2002)

have recently concentrated on a special case of Equation (2) when $\alpha = 0$ and $\beta = \lambda$. We call this the *Sine model* whose pdf is given by

$$f_{\mathrm{s}}(\phi, \psi) = C \exp\{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \lambda \sin(\phi - \mu) \sin(\psi - \nu)\} \tag{3}$$

where the normalizing constant is given by

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

and $I_r(\kappa)$ denotes the modified Bessels function of the first kind and order r.

Another such a sub-model of Equation (2) with similar characteristics is when $\alpha = \beta = -\kappa_3$. We will call this the *Cosine model* and its pdf is given by

$$f_{\mathrm{c}}(\phi, \psi) = \{c(\kappa_1, \kappa_2, \kappa_3)\} \exp\{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) - \kappa_3 \cos(\phi - \mu - \psi + \nu)\} \tag{4}$$

for $\kappa_1 \geq \kappa_3 \geq 0, \kappa_2 \geq \kappa_3 \geq 0$. Here the normalizing constant is given by

$$c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \{I_0(\kappa_1) I_0(\kappa_2) I_0(\kappa_3) + 2 \sum_{p=1}^{\infty} I_p(\kappa_1) I_p(\kappa_2) I_p(\kappa_3)\}.$$

The marginal probability density of ψ for the Cosine model (4) is given by

$$f_{\mathrm{c}}(\psi) = c(\kappa_1, \kappa_2, \kappa_3)\, 2\pi I_0(\kappa_{13}(\psi - \nu)) \exp\{\kappa_2\ \cos(\psi - \nu)\} \tag{5}$$

where $\kappa_{13}(\psi)^2 = \kappa_1^2 + \kappa_3^2 - 2\kappa_1\kappa_3 \cos\psi$. The marginal probability density of ϕ is given by an analogous expression. This distribution is symmetric about ν and is approximately a von Mises distribution for small values of κ_3. For $\kappa_1 = \kappa_2 = \kappa_3 = 0$, the distribution is uniform. For $\kappa_1 = \kappa_2 = 0$, the distribution is von Mises. Further, the conditional density $f_{\mathrm{c}}(\phi|\psi)$ is von Mises $\mathrm{M}(\psi_\nu, \kappa_{13}(\psi))$, where $\tan\psi_\nu = -\kappa_3 \sin(\psi - \nu)/(\kappa_1 - \kappa_3 \cos(\psi - \nu))$. An analogous expression holds for $f_{\mathrm{c}}(\psi|\phi)$.

Singh *et al.* (2002) have given the marginal and conditional distributions for the Sine model which we state here for comparison. The marginal density of ϕ (when $\mu = 0$) is proportional to

$$f_{\mathrm{s}}(\phi) = I_0(\kappa_{2\lambda}(\phi)) \exp\{\kappa_1 \cos\phi\}, \tag{6}$$

where $\kappa_{2\lambda}(\phi)^2 = \kappa_2^2 + \lambda^2 \sin^2\phi$. This is symmetric about $\mu = 0$ but not von Mises. Using Equations (3) and (6) we have the conditional probability density function of Ψ given $\Phi = \phi$ as $\mathrm{M}(\phi_\mu, \kappa_{2\lambda}(\phi))$ where $\tan\phi_\mu = (\lambda/\kappa_2)\sin\phi$.

3 Properties of Bivariate Circular Models

3.1 Bimodality Conditions

In this section, we state some key results for the sine and cosine models. The proofs are given in Web Appendix A.

Note that the joint density function (Φ, Ψ) in equation (4) (with $\mu = \nu = 0$) is also a bivariate von Mises for $f_{\phi,-\psi}$ and $f_{-\phi,\psi}$. Also $f_{\phi,0}$ and $f_{0,\psi}$. are univariate von Mises, and for the Sine model we have similar results.

Theorem 1 *For the Cosine density, when κ_1 and κ_2 are large, the random variable (Φ, Ψ) is approximately bivariate normal distributed if and only if $\kappa_3 \leq \kappa_1\kappa_2/(\kappa_1 + \kappa_2)$.*

Remark Derivations for bivariate normal approximations for the Sine model yields similar results and can be found in Singh *et al.* (2002).

Theorem 2 *The joint density function (Φ, Ψ) of the Cosine model in Equation (4) is unimodal if $\kappa_3 < \kappa_1\kappa_2/(\kappa_1+\kappa_2)$ and is bimodal if $\kappa_3 > \kappa_1\kappa_2/(\kappa_1+\kappa_2)$ when $\kappa_1 > \kappa_3 > 0, \quad \kappa_2 > \kappa_3 > 0$.*

Theorem 3 *The joint density function of the Sine model in Equation (3) is unimodal if $\kappa_1\kappa_2 > \lambda^2$ and is bimodal if $\kappa_1\kappa_2 < \lambda^2$, when $\kappa_1 > 0, \kappa_2 > 0$ and $-\infty < \lambda < \infty$.*

Remark We note that Singh *et al.* only discuss bimodality for the marginal distributions of the Sine density, and not the joint distribution as given here.

The marginal distribution of ψ for the Cosine model is given by Equation (5). It is easy to construct an example ($\kappa_1 = 1.5$, $\kappa_2 = 1.7$ and $\kappa_3 = 1.3$) for which the marginal distribution is unimodal even though the bivariate density is bimodal for the same κ values. This feature may be surprising at first: that, for a given set of parameters such that the Cosine density is bimodal, the corresponding marginal density may be unimodal. The following theorem describes the conditions under which this occurs.

Theorem 4 *For the Cosine model (4) with $\kappa_3 \neq 0$, the marginal distribution of Φ is symmetric around $\psi = \mu$ and unimodal (respectively bimodal) with mode at μ (respectively with the modes at $\mu - \phi^*$ and $\mu + \phi^*$) if and only if $A(|\kappa_1 - \kappa_3|) \leq$ (respectively $>$)$|\kappa_1 - \kappa_3|\kappa_2/(\kappa_1\kappa_3)$ where ϕ^* is given by the solution to $\kappa_1\kappa_3 A(\kappa_{13}(\phi))/\kappa_{13}(\phi) - \kappa_2 = 0$, and $A(\kappa) = I_1(\kappa)/I_0(\kappa)$.*

3.2 Choosing a Model

Analytical Comparisons

Clearly, a change in the mean values (μ and ν) will just shift the origin, so we will consider $\mu = \nu = 0$ in what follows. We note the following comparisons, which can all be obtained by considering the joint pdf and conditional distributions:[1]

[1] throughout this table, we use the term “expected” as shorthand for the directional mean

	Sine model	Cosine model
1	Changing the sign of λ simply reflects in an axis.	Changing the sign of small κ_3 give an approximate reflection, but for $\|\kappa_3\|$ large (relative to κ_1, κ_2), bimodality occurs only for positive κ_3.
2	The expected value of $\phi\|\psi = 0$ is always 0 – for any values of $\kappa_1, \kappa_2, \lambda$.	The expectation is zero only when the $\kappa_i, i = 1, 2, 3$ give a unimodal pdf.
3	The expected value of $\phi\|\psi$ is always constrained to lie in the range $[-\pi/2, \pi/2]$ – for any values of $\kappa_1, \kappa_2, \lambda$ and ψ.	For large (negative) κ_3 the expected value of $\phi\|\psi$ approximates ψ over the full range of $(-\pi, \pi)$.
4	A bimodal density occurs for large $\|\lambda\|$ (relative to κ_1, κ_2).	A bimodal distribution occurs only for large positive κ_3.
5	Transforming (ϕ, ψ) to $(\psi, -\phi)$ is equivalent to changing the sign of λ.	This transformation allows for rotations of the pdf (which cannot be achieved by changing κ_3).

The bivariate densities can be represented by contour plots which can be used to illustrate the above statements. However, the key features are more easily compared by plotting the log of the density and omitting the normalizing constant. In numerical comparisons we noted that for small $\lambda \approx \kappa_3$ the two models are very similar.

Overall, we could conclude that, if we allow for the possibility of transformations of the form (ϕ, ψ) to $(\psi, \pi + \phi)$ and (ϕ, ψ) to $(\psi, -\phi)$ then the Cosine model gives a richer set of possible contour plots (see point 5 above), and hence should be able to fit more closely a larger class of distributions.

Numerical Comparisons

Another approach for comparing the Sine and Cosine models is to examine their moments. Characteristic functions can be used to estimate the moments of the two distributions and we numerically computed the correlations for the Cosine and Sine models in order to study their behaviour in relation to the parameters κ_3 and λ. The parameter κ_3 for the Cosine model and λ for the Sine model were varied for fixed values of κ_1 and κ_2. Using some selected values of κ_1 and κ_2, we observed that the correlation between $\cos\phi$ and $\cos\psi$ and between $\sin\phi$ and $\sin\psi$ were seen to be (mostly) decreasing functions of κ_3 for the cosine model, whereas for the Sine model the correlation between $\sin\phi$ and $\sin\psi$ was a monotonic increasing function of λ, and between $\cos\phi$ and $\cos\psi$ was always non-negative and has a "U"-shaped relationship with λ. Thus we believe that the Cosine model may be better able to capture the correlation between ϕ and ψ as either correlation of $(\cos\phi, \cos\psi)$ or $(\sin\phi, \sin\psi)$. In the Sine model, λ does not measure the correlation of $(\cos\phi, \cos\psi)$, and his would seem to give the Cosine model an advantage over the Sine model.

It is also of interest to empirically explore the relationship between the bivariate normal distribution and the Cosine and Sine models when the data are highly concentrated. We estimate the correlation under the assumption of normality induced by high concentration. We compare this estimated correlation under the assumption of normality with correlation evaluated from the distributions by integration. Under high concentration, the approximate correlation between ϕ and ψ, for the Cosine model, is

$$\rho_c = \frac{-\kappa_3}{\sqrt{(\kappa_1 - \kappa_3)(\kappa_2 - \kappa_3)}}$$

and for the Sine model Singh *et al.* (2002) obtain $\rho_s = \lambda/\sqrt{\kappa_1\kappa_2}$. We compared the values of ρ_c and ρ_s with the correlation of $(\cos\phi, \cos\psi)$ and $(\sin\phi, \sin\psi)$ for various large (κ_1, κ_2). Overall, we found that the Cosine model has some advantages over the Sine model. Sometimes the Sine model can explain the presence of correlation only in the range of $(0, 0.4)$. In the Cosine model, the “estimated correlation” tracks the values of the parameter κ_3 very well for the entire range of values between $(-1, 1)$. For further details, see Subramaniam (2005).

4 Application: Mixture Models for Proteins

4.1 Protein Data

Protein structures can be determined to an atomic level by X-ray diffraction and neutron-diffraction studies of crystallized proteins, and more recently by nuclear magnetic resonance (NMR) spectroscopy of proteins in solution. In a protein, the backbone chain N-C$_\alpha$ and C$_\alpha$-C bonds are relatively free to rotate. These rotations are represented by the torsion angles ϕ and ψ, respectively.

A scatter planar representation of such torus data has come to be known as a *Ramchandran plot* (Ramachandran *at al.*, 1963), and a study of these angles through directional statistics methods has been one of our goals. Ramachandran used computer models of small polypeptides to systematically vary ϕ and ψ with the objective of finding stable conformations. For each conformation, the structure was examined for close contacts between atoms. Atoms were treated as hard spheres with dimensions corresponding to their van der Waals radii. Therefore, ϕ and ψ angles which cause spheres to collide correspond to sterically disallowed conformations of the polypeptide backbone.

We consider two datasets which correspond to these conformational angles from the proteins Malate dehydrogenase (7mdh in the protein database: http://www.rcsb.org/pdb) and Myoglobin (protein 101m). Myglobin is the smallest protein, consisting mainly of α-helices, and the first protein whose structure was determined (Bennett & Kendrew, 1952). Malate dehydrogenase is an “average” protein, with various motifs.

The Ramachandran plot and the circular histograms of the marginal variables for Malate dehydrogenase are shown in Figure 1. These clearly indicate that a mixture distribution may be appropriate since there are obviously a number of distinct clusters corresponding to the secondary structure of the protein. The marginals show departure from the von Mises distribution which only has one mode.

[Figure 1 about here.]

4.2 Mixture Models

We can use the method of maximum likelihood to estimate the parameters of a cosine density. The log-likelihood function for Equation (4) is

$$\sum_{i=1}^{n} (\kappa_1 \cos(\phi_i - \mu) + \kappa_2 \cos(\psi_i - \nu) - \kappa_3 \cos(\phi_i - \psi_i + \nu - \mu)) + n \log(c(\kappa_1, \kappa_2, \kappa_3)). \quad (7)$$

and we can obtain starting values under the assumption that the marginal distributions are von Mises. Then, for example, starting values for μ are given by

$$\overline{\phi} = \begin{cases} \tan^{-1}(\overline{S}/\overline{C}) & \text{if } \overline{C} \geq 0; \\ \tan^{-1}(\overline{S}/\overline{C}) + \pi & \text{if } \overline{C} < 0; \end{cases}$$

where $\overline{C}$ and $\overline{S}$ are the means of the $\cos(\phi)$ and $\sin(\phi)$, respectively, and starting values for ν are similarly found from the ψs. For $\hat{\kappa}_1$ an approximate solution due to Dobson (1978) is used. The starting values are obtained for ν and $\hat{\kappa}_2$ in the same way. For $\hat{\kappa}_3$ we use the mean of $(\hat{\kappa}_1, \hat{\kappa}_2)$. Our optimization program for maximizing the (log-) likelihood was tried and tested using data simulated from the bivariate Cosine distribution. For details of the simulation method, see Web Appendix B.

However, it is clear from Figure 1 that a single Cosine model will not fit these data — even though a bimodal distribution can sometimes be obtained. The plot suggests $K \geq 3$

components in a mixture model, which can be parameterized by:

$$f_M(\phi, \psi) = \sum_{j=1}^{K} \pi_j f_j(\phi, \psi) \tag{8}$$

where f_j denotes a Cosine density with parameters $\theta_j = (\kappa_{jk}, k = 1, 2, 3, \mu_j, \nu_j),\ j = 1, \ldots, K$, and $\pi_1, \ldots, \pi_K$ are the mixing proportions (with $\sum \pi_i = 1$).

4.3 EM Algorithm

We will investigate the use of the EM algorithm (McLachlan & Krishnan, 1997, pp. 71–72) to fit (8). As usual, there are several steps which are iterated to convergence:

1. Estimation of membership probabilities using

   (a) $p_{ij} = \pi_j f_j(\phi_i, \psi_i)$ for $i = 1, \ldots, n, j = 1, \ldots, K$

   (b) Normalization $p_{ij} = p_{ij} / \sum_j p_{ij}$ for $i = 1, \ldots, n$

2. Use maximum likelihood to find the $\theta_j = (\kappa_{j1}, \kappa_{j2}, \kappa_{j3}, \mu_j, \nu_j)$ which maximizes the weighted likelihood function

   $$\prod_{i=1}^{n} p_{ij} f_j(\phi_i, \psi_i)$$

   for $j = 1, \ldots, K$.

3. Obtain the mixing proportions $\pi_j = \sum_i p_{ij}$

However, in our use of the Cosine density we have increased the ability to model the data by also considering transformations of all the data the form $(\phi, \psi) \to (\psi, -\phi)$ for estimating the parameters of each mixture component. In practice this means checking the likelihood (for each j) at step 2., and choosing the rotation which gives the larger value.

It is well-known that the EM algorithm can get stuck in local solutions, and there can also be a problem with singularities in which one of the components consists of only a single observation. In our implementation, we tried several starting values, and chose the best final solution, excluding any solutions with very high concentrations (large κ).

The choice of the number of components K will obviously affect the final likelihood. Adding another component will increase the number of parameters by 6, and this was done incrementally (starting with $K = 1$) until the first minimum of AIC, and it is this value of K which is then reported. After convergence to the final solution we can note the mixing proportions $\pi_1, \ldots, \pi_K$, the parameter estimates $\theta_j, j = 1, \ldots, K$ and the membership probabilities p_{ij}. Then, for each observation (ϕ_i, ψ_i) we can assign it to the group which maximizes p_{ij}.

4.4 Results

The EM algorithm has been used to fit mixtures of the cosine model to both Malate dehydrogenase and Myoglobin conformational angles.

For Myoglobin, the algorithm selected $K = 3$ components, all of which used the rotated parameterization. The contour plots of the log densities are shown in Figure 3, and the parameter estimates are given in Table 1. We note that the first and second components have fit a cosine model which results in a bimodal density, and in this case Theorem 4 can be used to locate the modes exactly. Further, we note that the third component (which contains nearly $3/4$ of the observations) has formed a very tight cluster about its mean. [As a rough guide, κ values around 5 are considered as “moderate concentration”, and values over 10 are considered as “high concentration”.] Myoglobin is known to contain nearly all helices, and this dominance of one of the components is consistent with this fact.

[Figure 2 about here.]

[Table 1 about here.]

Historically, describing protein secondary structure has been somewhat subjective, but the DSSP program (Kabsch & Sander, 1983; http://swift.cmbi.kun.nl/swift/dssp) was designed to standardize the assignment from protein database co-ordinates to secondary structures. This program will convert a protein database file to (ϕ, ψ) angles and assigns each pair of angles to one of:

"loop or irregular" (for which no label is given, *i.e.* blank)

B residue in isolated beta-bridge

E extended strand, participates in beta ladder

G 3-helix (3/10 helix)

H alpha helix

I 5 helix (pi helix)

S bend

T hydrogen bonded turn

It is of interest to compare the clusters formed by our 3-mixture model with the classifications given by DSSP and these are shown in Table 2. Almost all of the "helix" (H) angles are contained in the third cluster, which is the most highly concentrated group; indeed the third cluster (which has very high concentration) contains only angles labelled H (alpha helix), or G (3/10 helix), and one hydrogen bonded turn. Figure 2 shows the Ramachandran plot for Myoglobin with the DSSP classifications used as labels.

[Table 2 about here.]

The second protein for which we present results is Malate dehydrogenase, which is somewhat larger (343 angles), with several motifs, including β-sheets. The DSSP classifications are shown in a Ramachandran plot in Figure 1. In this case, the EM algorithm was used to find solutions for $K = 1, 2, \ldots, 6$ in which AIC selected $K = 5$ components, of which three used the parameter rotations. The contour plots of the log densities are shown in Figure 3, and the parameter estimates are given in Table 1. Again, we note that one of the components (4) has formed a very tight cluster about its mean, and that component 5 has a bimodal density . Comparison of our cluster labels with those obtained from DSSP is given in Table 2. In this table we see that cluster 4 is almost all alpha helix, but that many helices are also in cluster 3.

For both proteins we observe that the mean values of the components in the mixture models are consistent with the traditional partitions given to the Ramachandran plots based on an empirical distribution analysis (e.g. Branden & Tooze, 1998, p. 9). Moreover, we note that the model selection has fit more components to Malate dehydrogenase (which contains more types of secondary structures) than for Myoglobin (which does not contain any β-sheets (DSSP class E)).

[Figure 3 about here.]

5 Discussion

There are various future directions which are possible. For example, we could take account of any serial correlation between the conformational angles ψ_i and ψ_{i+1}, or between ϕ_i and ϕ_{i+1}, since we may not have i.i.d. data. This serial correlation could be modelled (for example by some sort of Markov Chain) or a simpler solution would be “thinning” in which only every fourth (say) observation was retained. Serial correlation models could also be used to adjust the posterior probabilities of membership p_{ij}, so that isolated angle membership was made less likely than an independence assumption would permit. These are subjects of ongoing research. Note also, the possibility of allowing for two additional classifications: a class “doubt” if all $p_{ij} < t$ for some threshold t, and a class “uncertain” if the two largest probabilities are almost equal.

Clearly other submodels are also possible such as the bivariate wrapped normal, Cauchy, but these are not members of the exponential family. However, when applied to the protein data, it is clear that several mixture components will be required. In our case the number of parameters for K mixtures is $6K - 1$, so for the larger 5 component model we have 29 parameters. However, we think that there is still good interpretability – and standard errors are also available. This contrasts with the model of Pertsemlidis *et al.* (2005) who used about 100 parameters which cannot easily be intrepreted. The final mixture model and corresponding set of clusters is objective, and has much similarity with both initial subjective classifications, and the output from DSSP.

Extending to more than 2 dimensions is clearly possible both for the Sine model and Cosine model, but is somewhat easier to express for the Sine model. Further work in higher dimensions is in progress. Finally, we note that the methods of this paper can be used in other applications with bivariate angular data, for example wind directions (Fisher, 1993).

Supplementary Materials

Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgements

We would like to thank Dave Westhead, John Kent, Thomas Hamelryck and Andy Neuwald for useful discussion and help with the data examples. We would also like to thank an anonymous referee, the assosicate editor, and the editor for helpful suggestions which have greatly helped to improve the paper.

References

Bennett, J.M. & Kendrew, J.C. (1952). The computation of Fourier syntheses with a digital electronic calculating machine. *Acta Crystallographica*, **5**, 109–116.

Branden, C. and Tooze, J. (1998) *Introduction to Protein Structure*. Garland.

Dobson, A. J. (1978). Simple approximations to the von Mises concentration statistic. *Applied Statistics*, **27**, 345–346.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

Hamelryck, T. (2005). MOCAPY `https://sourceforge.net/projects/mocapy/`

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. **22**, 2577–637.

Mardia, K. V. (1975). Statistics of directional data (with discussion). *J. Royal Statistical Society*, **B37**, 349–393.

Mardia, K.V. and Jupp, P.E. (1999) *Directional Statistics*. Wiley.

McLachlan, G. F. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley, New York.

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.

Ramachandran, G.N., Ramakrishnan, C. & 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*, **A17**, 461–483.

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

Subramaniam, G. (2005). *von Mises Distributions with Applications in Speech Data*. PhD thesis, Department of Statistics, University of Leeds, Leeds U.K.

List of Figures

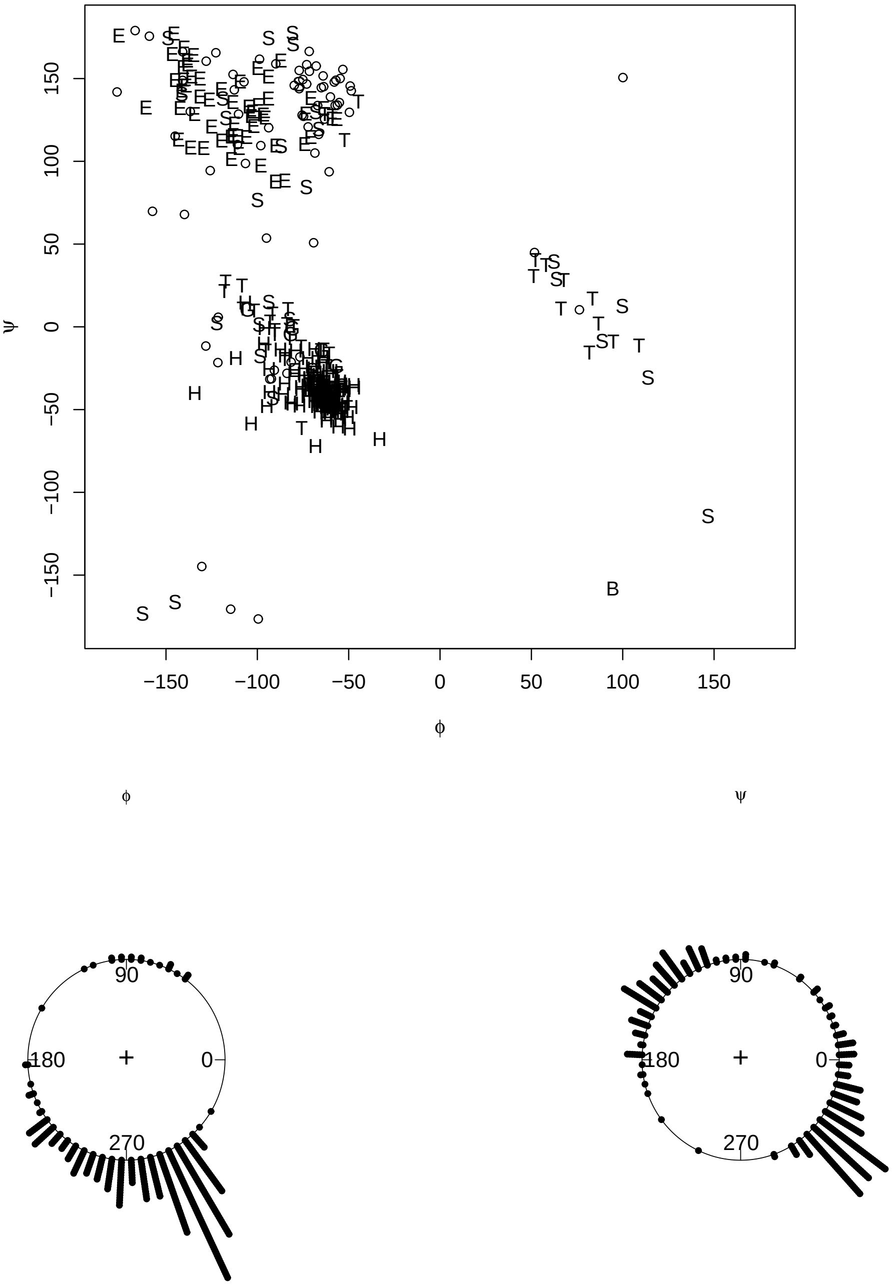

Figure 1: Ramachandran plot of Malate dehydrogenase (top), and Circular Plots of ϕ, ψ (bottom). The symbols used in the plot correspond to the DSSP classifications; see Section 4.4

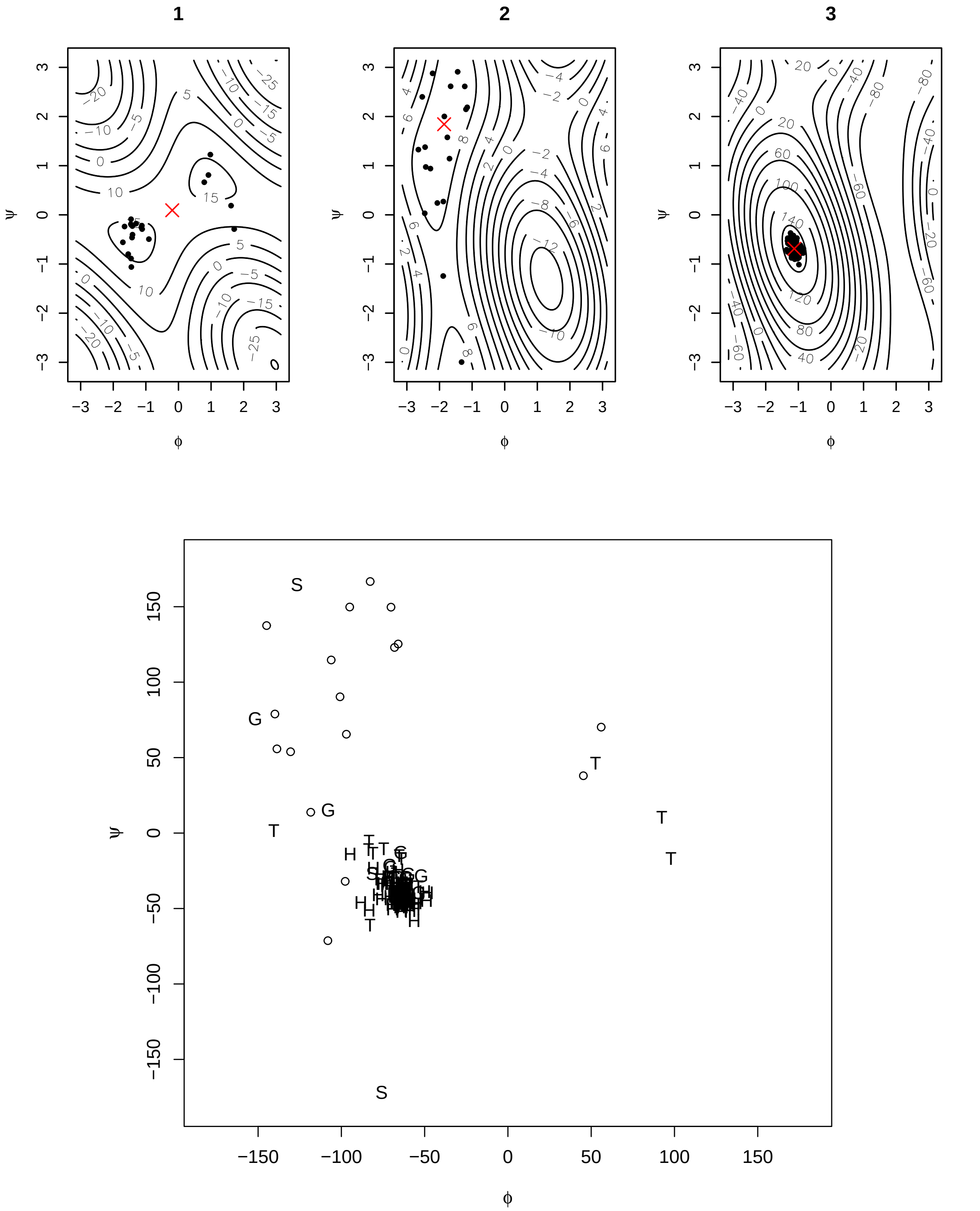

Figure 2: Mixture model for Myoglobin. Top three panels correspond to the three components, with contour plots of the log density, and points allocated to the most probable mixture. The location × (sometimes obscurred in a cluster of points) marks the mean (μ, ν) for each mixture. Bottom: Ramachandran plot of Myoglobin, with DSSP classifications. (The unlabelled points, given by open circles, are "blank" in DSSP.)

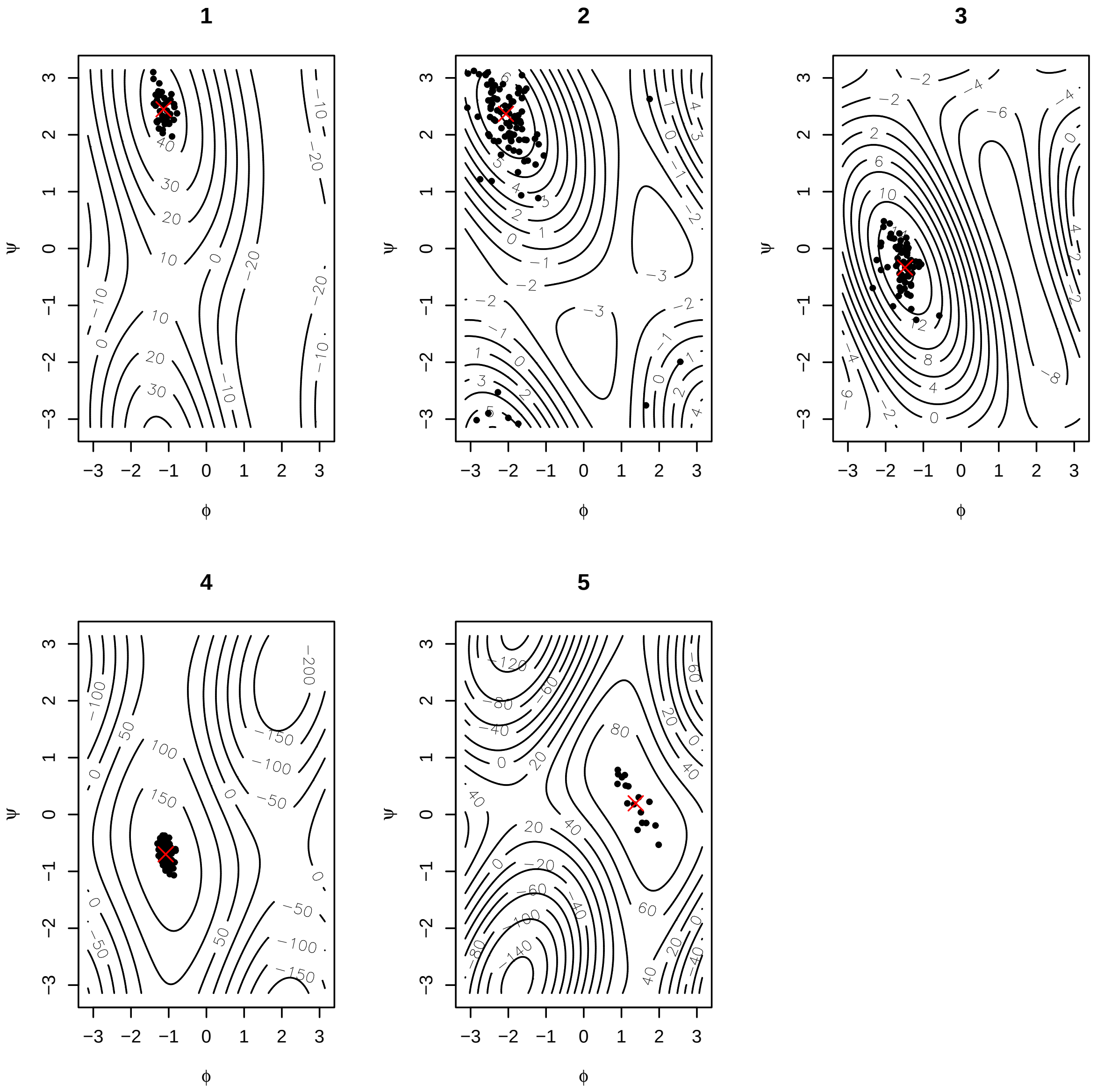

Figure 3: Mixture model for Malate dehydrogenase. Five panels correspond to the five components, with contour plots of the log density, and points allocated to the most probable mixture. The location $\times$ (sometimes obscurred in a cluster of points) marks the mean (μ, ν) for each mixture.

List of Tables

Myoglobin

component		κ_1	κ_2	κ_3	μ	ν	π	R	modes
1	est.	8.459	13.544	8.081	-0.191	0.093	0.138	Y	B
	SE	1.881	3.243	1.819	0.129	0.073	0.028		
2	est.	8.340	3.989	2.872	-1.858	1.843	0.124	Y	B
	SE	2.546	1.340	1.230	0.094	0.220	0.027		
3	est.	89.075	39.679	-38.603	-1.124	-0.689	0.738	Y	U
	SE	14.700	10.181	10.108	0.009	0.012	0.036		

Malate dehydrogenase

component		κ_1	κ_2	κ_3	μ	ν	π	R	modes
1	est.	27.759	9.491	-8.669	-1.136	2.449	0.115	Y	U
	SE	7.283	4.407	4.325	0.028	0.040	0.017		
2	est.	2.495	2.434	-2.119	-2.063	2.367	0.281	Y	U
	SE	0.509	0.504	0.484	0.058	0.059	0.024		
3	est.	7.075	3.741	-5.405	-1.486	-0.335	0.219	Y	U
	SE	1.529	1.215	1.354	0.039	0.046	0.022		
4	est.	134.897	83.831	28.405	-1.08	-0.699	0.336	N	U
	SE	18.960	12.900	7.769	0.01	0.013	0.025		
5	est.	70.840	62.464	34.960	1.378	0.197	0.050	N	B
	SE	22.557	19.917	11.525	0.061	0.069	0.012		

Table 1: Estimates (μ and ν are given in radians) and standard errors for the mixture components fitted to Myoglobin (top; three components) Malate dehydrogenase (bottom; 5 components). The final two columns indicate: $R = Y$ that the component parameters are fitted to rotated data (otherwise $R = N$) and modes $= B$ that the density is bimodal (otherwsie unimodal (U)).

Myoglobin

Gp		G	H	S	T
1	3	2	4	1	10
2	14	2	0	2	1
3	0	8	104	0	1

Malate dehydrogenase

Gp		B	E	G	H	S	T
1	29	0	5	0	0	5	4
2	30	1	51	0	0	11	0
3	9	0	1	4	26	6	20
4	0	0	0	2	113	1	8
5	2	0	0	0	0	5	10

Table 2: Cross-classification based on the labels from the DSSP program (columns) and the clusters obtained from the K-mixture cosine distribution model. Top: $K = 3$ for Myoglobin. Bottom: $K = 5$ for Malate dehydrogenase.