Circular time series with application to protein conformations

Gareth Hughes*, Kanti V. Mardia and Charles C. Taylor

University of Leeds

1 Introduction

The 3-dimensional conformation of proteins is an active research area that has received much attention. Much of the statistical modelling of these conformations has been investigated with bivariate circular models (see for example Singh *et al.* (2002), Mardia *et al.* (2003)). In this poster we consider three first order autoregressive time series models for univariate directional data which have the potential for extension to higher orders and, ultimately, bivariate directional time series models.

2 ϕ and ψ conformational angles

The backbone of a polypeptide chain comprises a sequence of atoms

$$[N_1{-}C_1^\alpha{-}C_1]{-}[N_2{-}C_2^\alpha{-}C_2]{-}\ldots{-}[N_p{-}C_p^\alpha{-}C_p],$$

in which the lengths of the bonds between any two successive atoms and the angle between any three successive atoms are approximately constant. The degrees of freedom of the polypeptide chain therefore involve angles within chains of 4 atoms. Since there are 3 different atoms in the backbone (N, C and C^α), there are 3 angles to be considered. In each case, the angle concerned is that made by the fourth atom with the plane defined by the previous three. A zero direction and positive orientation are determined by the right-hand rule, by pointing the right-hand thumb in the direction of the bond linking atoms 2 and 3, placing the palm on the plane of atoms 1-3 (zero direction), and curling the fingers (positive orientation). Angles are measured between $-\pi$ and π.

ϕ_i = angle of rotation of C_i atom around N_i–C_i^α bond, with reference to the C_{i-1}–N_i–C_i^α plane, $i = 2, \cdots, p$.

ψ_i = angle of rotation of N_{i+1} atom around C_i^α–C_i bond, with reference to the N_i–C_i^α–C_i plane, $i = 1, \cdots, p-1$.

ω_i = angle of rotation of C_{i+1}^α atom around the peptide bond C_i–N_{i+1}, with reference to the C_i^α–C_i–N_{i+1} plane, $i = 2, \cdots, p$.

Note from the above that ϕ_1 and ψ_p are undefined. The angle ω is restricted to be around zero and is therefore not considered in the following analysis.

3 Conditional probability models

A circular variable x (i.e. an angle) is said to follow a von Mises distribution with mean direction μ and concentration parameter κ if its pdf is given by

$$f(x) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa \cos(x - \mu)\} \tag{3.1}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero. In this case we write $x \sim M(\mu, \kappa)$.

We now outline three conditional probability models, two of which are determined from a joint probability distribution.

3.1 Möbius model

Downs and Mardia (2002) constructed a circular regression model in which a dependent angular variable v is modelled conditionally on an independent angular variable u. This model can be adapted to a circular time series context by replacing v with θ_t and u with θ_{t-1}, so that

$$\Theta_t|(\Theta_{t-1} = \theta_{t-1}) \sim M\big(\alpha_1 + 2\tan^{-1}\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha_1)\}, \kappa\big), \qquad t = 2, \cdots, n, \tag{3.2}$$

in which $\omega \in [-1, 1]$ is a slope parameter and $\alpha_1 \in (-\pi, \pi]$ an angular location parameter.

3.2 Sine and Cosine models

The so–called Sine and Cosine models arise as special cases of a set of bivariate models considered by Rivest (1987), namely

$$\begin{aligned} f_1(\theta, \phi) = C \exp\{&\kappa_1 \cos(\theta - \mu) + \kappa_2 \cos(\phi - \nu) + \\ &+ \alpha \cos(\theta - \mu)\cos(\phi - \nu) + \beta \sin(\theta - \mu)\sin(\phi - \nu)\}. \end{aligned} \tag{3.3}$$

This is itself a submodel of a class considered by Mardia (1975). With $\theta = \theta_{t-1}$ and $\phi = \theta_t$, the parameterization $\alpha = 0$, $\beta = \lambda$, $\nu = \mu = \alpha_2$, say, and $\kappa_1 = \kappa_2 = \kappa$, say, gives the conditional distribution of Θ_t given $\Theta_{t-1} = \theta_{t-1}$ corresponding to (3.3) as

$$\Theta_t|(\Theta_{t-1} = \theta_{t-1}) \sim M(\alpha_2 + \tan^{-1}\{\kappa^{-1}\lambda \sin(\theta_{t-1} - \alpha_2)\}, \kappa^*), \qquad t = 2, \cdots, n, \tag{3.4}$$

where $\kappa^* = \sqrt{[\kappa_2^2 + \lambda^2 \sin^2(\theta_{t-1} - \alpha_2)]}$. Equation (3.4) is the conditional distribution of the Sine model.

The parametrization $\alpha = \beta = a$, $\kappa_1 = \kappa_2 = b$, $\mu = \nu = \alpha_3$ (say) gives the conditional distribution of Θ_t given $\Theta_{t-1} = \theta_{t-1}$ corresponding to the joint density (3.3) as

$$\Theta_t|(\Theta_{t-1} = \theta_{t-1}) \sim M(\tan^{-1}(a \sin\theta_{t-1} + b \sin\alpha_3, a\cos\theta_{t-1} + b\cos\alpha_3), \kappa_t), \tag{3.5}$$

where the time varying concentration $\kappa_t = \sqrt{[a^2 + b^2 + 2ab\cos(\theta_{t-1} - \alpha_3)]}$ and where $[\tan^{-1}(d, c)] \in (-\pi, \pi]$ is the angle between the x–axis and the vector (c, d). Equation (3.5) is the conditional distribution of the Cosine model.

For all three models (3.2), (3.4) and (3.5), all θ angles are taken to be in the half-open interval $(-\pi, \pi]$.

4 Model properties

To investigate the properties of the Möbius, Sine and Cosine models, the behaviour of μ_t given $\Theta_{t-1} = \theta_{t-1}$, where μ_t is the mean direction of Θ_t given $\Theta_{t-1} = \theta_{t-1}$, is investigated for each of the three models for values of θ_{t-1} ranging from $-\pi$ to π.

Figure 1 shows a graphical representation of this behaviour, based on the parameter values displayed in the plots. The appearance of the plots is clearly dependent on the value of the

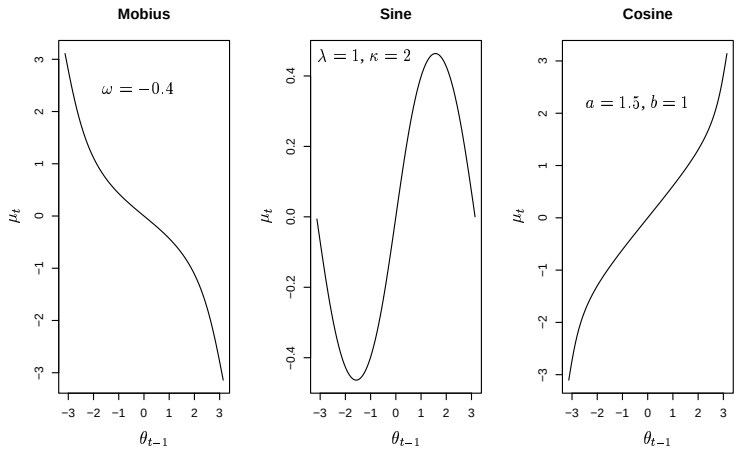

Figure 1: Plots of μ_t, the mean direction of θ_t given $\Theta_{t-1} = \theta_{t-1}$, for the Möbius, Sine and Cosine models. $\alpha_i = 0$ for $i = 1, 2, 3$; other parameter values are shown in the plots.

parameters, and we give a discussion of the parameter constraints under which certain features occur.

It is concluded that the Sine model (3.4) is inappropriate since the value of μ_t given $\Theta_{t-1} = \theta_{t-1}$ is the same as the value of μ_t given $\Theta_{t-1} = \pi - \theta_{t-1}$, as can be seen from Figure 1. This is in the case $\alpha_2 = 0$, although the symmetry remains (shifts) if α_2 changes.

Another undesirable property, or rather a property that one would not expect to observe from an angular data set, that can be observed for the Sine model in Figure 1 is that, for $|\theta_{t-1}| > \pi/2$, the further θ_{t-1} is from the overall mean ($\alpha_2 = 0$), the closer μ_t is to this value. For these reasons the Sine model was not applied to the protein data.

5 Application to protein data

Based on the conclusions of the discussion of the model properties, suitable models are fitted to the ϕ and ψ angles (individually) of a polypeptide chain, and the results and model adequacy discussed. MLE's of the parameters are obtained by numerical maximisation of the conditional likelihood function

$$f(\theta_2, \cdots, \theta_n|\theta_1) = f(\theta_2|\theta_1) f(\theta_3|\theta_2) \cdots f(\theta_n|\theta_{n-1}) \tag{5.6}$$

Figure 2 shows a "time series" plot of 343 consecutive ϕ angles from a polypeptide chain, along with a kernel density estimate based on these data. The latter has been centred for clarity, which explains the shift from $(-\pi, \pi]$ on the x–axis.

6 Conclusions and further work

The main conclusions from fitting the Möbius and Cosine models are that the fit is inadequate. These conclusions are based on a comparison with the protein data, for which MLE's are calculated, and data generated using these maximum likelihood estimates. Of course, as in a linear time series context, an $AR(1)$ will not be adequate for every data set, and increasing the order of the model will improve its fit. The next stage of this work, therefore, is to extend the models to a higher order, after which bivariate models will be sought.

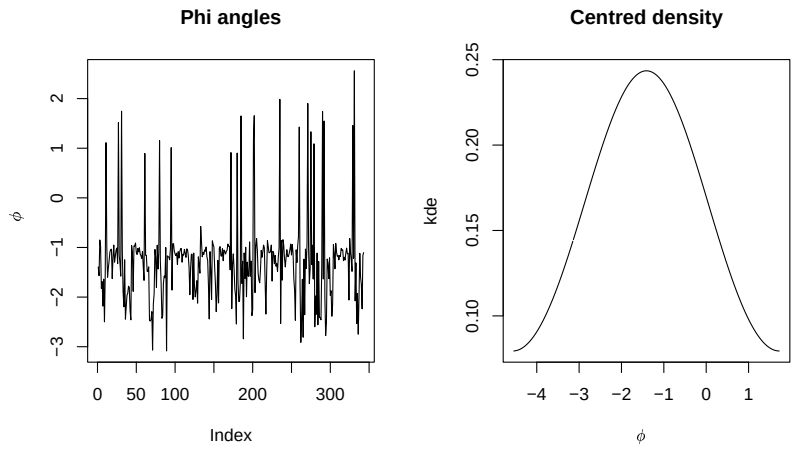

Figure 2: A sequence of 343 consecutive ϕ angles from a polypeptide chain, and a centred kernel density estimate based on these angles.

Acknowledgement: We would like to thank David R. Westhead (University of Leeds) for his guidance with the bioinformatic aspects of this work.

References

Downs, T. D. and Mardia, K. V. (2002). Circular regression. *Biometrika*, **3**, 89, 683–697.

Mardia, K. V. (1975). Statistics of Directional Data. *J. R. Statist. Soc. B*, **3**, 37, 349–393.

Mardia, K. V., Taylor, C. C. and Subramaniam, M (2003). Application of circular distributions to conformational angles in protein. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 149–152. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon. Leeds University Press.

Rivest, L. P. (1987). A distribution for dependent unit vectors. *Communications in Statistics A*, **17**, 461–483.

Singh, H., Hnizdo, V. and Demchuk, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika*, **3**, 89, 719–723.