

Directional data on the torus, with applications to protein structure

Charles C. Taylor¹

School of Mathematics, e-mail: c.c.taylor@leeds.ac.uk

Abstract: For data in which the both the response, and all the explanatory variables are angles (measured in $[-\pi, \pi)$) we propose a new regression model which makes use of a more general transformation. Properties of the simple case (of one explanatory variable) are stated, and comparisons made with previous models. Extensions to “polynomial-type” models and multiple regression are outlined, and an example using conformational angles from proteins is given.

Keywords: Angular, Circular, Möbius transformation, von Mises

1. Introduction

Directional data, or circular data can be expressed as angles in $[-\pi, \pi)$ measured clockwise, or anti-clockwise from some specified *origin*. Models for, and analysis of univariate circular data are well-developed (Mardia & Jupp, 2000; Jammalamadaka & SenGupta, 2001) but methods for multivariate data (except under high concentration) need further research. In this paper, we consider the challenge of circular-circular polynomial, and multiple regression in which the data lie on a d -dimensional torus.

Initially we focus on the case of bivariate circular data ($d = 2$) in which y is a response variable, and x an explanatory variable. We propose a regression model in which the conditional distribution is described by

$$Y \mid x \sim \text{vM}(\text{atan2}(a \sin x, b \cos x + 1), \kappa). \quad (1)$$

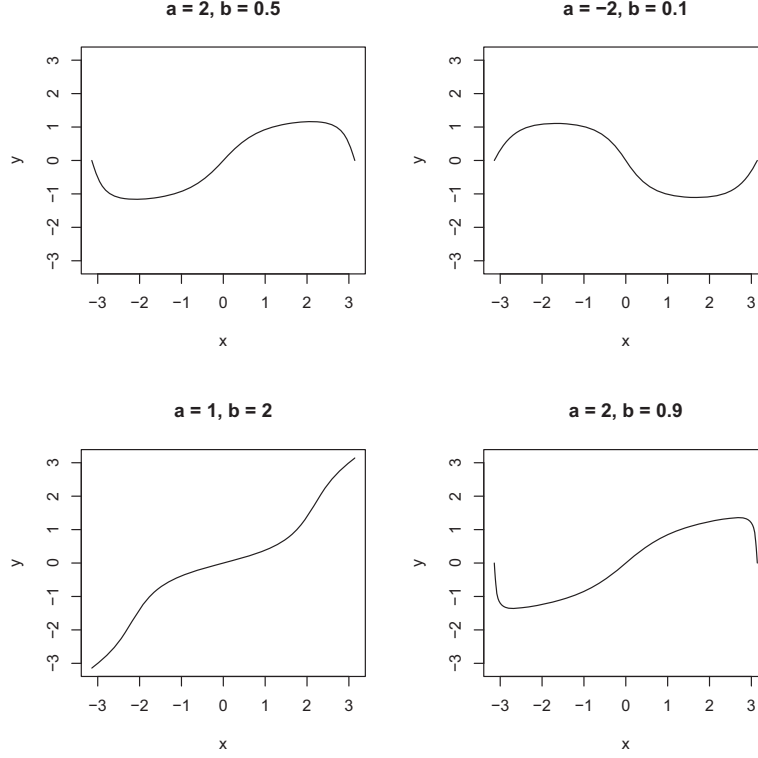
where a and b are parameters of the model, and the function $\text{atan2}(w, v)$ gives the angle between the x -axis and the vector from the origin to (v, w) . The conditional error distribution is von Mises with concentration κ , i.e. $f(\varepsilon \mid x) = (2\pi I_0(\kappa))^{-1} \exp\{\kappa \cos(\varepsilon)\}$ with $I_\nu(\kappa)$ the modified Bessel function of the first kind and order ν . Note that model (1) assumes that the regression curve passes through the origin. More generally we could write $Y \mid x \sim \text{vM}(\mu_y + \text{atan2}(a \sin(x - \mu_x), b \cos(x - \mu_x) + 1))$ but for simplicity we can initially centre the data by subtracting the circular means from each variable.

Figure 1 shows some examples of regression lines for different values of a, b . It is easy to show that, if $|b| = 1$ the regression curve is not continuous at $x = \pm\pi$, and, in general, that the gradient at the origin is $a/(b + 1)$, and at $x = \pm\pi$ the gradient is $a/(b - 1)$. For $b > 1$ the curve passes through $(\pm\pi, (2I(a > 0) - 1) \pm \pi)$ and for $b < 1$, $y = 0$ when $x = \pm\pi$. There are turning points at $(x, y) = (\text{atan2}(\pm\sqrt{1 - b^2}, -b), \text{atan2}(\pm a\sqrt{1 - b^2}, 1 - b^2))$ which are real only if $|b| < 1$. Finally, for $b < -1$ the curve does not pass through the origin, and we note that values of (a, b, μ_x) are equivalent to $(-a, -b, \mu_x + \pi)$.

¹Address for correspondence: School of Mathematics, University of Leeds, Leeds LS2 9JT, UK

We now describe the relationship of our model to previous work, and then briefly outline estimation and inference of the parameters. We then generalize the model to polynomial forms, and multiple regression. We end with an example to conformational angles in protein structure.

Figure 1: Examples of model (1) for various values of a, b



2. Relationship to other Models

Natural generalizations of the von Mises distribution to higher dimensions are given by Hughes (2007). In the bivariate case we have the *Sine* and *Cosine* models given by (in the centred case)

$$\begin{aligned} f_{\mathbf{S}}(x, y) &\propto \exp\{k_1 \cos x + k_2 \cos y + \lambda \sin x \sin y\} \\ f_{\mathbf{C}}(x, y) &\propto \exp\{k_1 \cos x + k_2 \cos y - k_3 \cos(x - y)\} \end{aligned}$$

respectively. This leads to the conditional distributions:

$$\begin{aligned} f_{\mathbf{S}}(y | x) &= (2\pi I_0(k_{\mathbf{S}}(x)))^{-1} \exp\{k_{\mathbf{S}}(x) \cos(y - \text{atan2}(\lambda \sin x, k_2))\} \\ f_{\mathbf{C}}(y | x) &= (2\pi I_0(k_{\mathbf{C}}(x)))^{-1} \exp\{k_{\mathbf{C}}(x) \cos(y - \text{atan2}(k_3 \sin x, k_3 \cos x - k_2))\} \end{aligned}$$

where $k_{\mathbf{S}}(x)^2 = k_2^2 + \lambda^2 \sin^2 x$ and $k_{\mathbf{C}}(x)^2 = k_2^2 + k_3^2 - 2k_2k_3 \cos(x)$, and both of which are von Mises. Substituting $k_1 = k_2$ it can be seen that these are special cases of our model (1). The filtering approach suggested by Kent & Mardia (2002) leads to the same conditional distribution as the Cosine model.

Downs and Mardia (2002) used a Möbius transformation to obtain the (centred) model $y | x \sim \text{vM}(2\text{atan2}(\omega \sin x/2, \omega \cos x/2), \kappa)$ which is similar to (1) with $a, b, \rightarrow \infty$ and $\omega = a/b$. The Möbius transformation was also considered (using a different representation) by Kato *et al.* (2008) but with a wrapped Cauchy distribution for the errors.

3. Estimation and Inference

Given data $(x_1, y_1), \dots, (x_n, y_n)$, then the MLEs of a and b satisfy

$$\begin{aligned} \sum \frac{\sin x_i \sin y_i}{v_i} \left(1 - \frac{a^2 \sin^2 x_i}{v_i^2} \right) &= 0 \\ \sum \frac{\cos x_i \cos y_i}{v_i} \left(1 - \frac{(b \cos x_i + 1)^2}{v_i^2} \right) &= 0 \end{aligned}$$

where $v_i^2 = a^2 \sin^2 x_i + (b \cos x_i + 1)^2$, and the estimate of κ is then given by

$$A_1^{-1}(\kappa) = \frac{1}{n} \sum \cos y_i \frac{\hat{b} \cos x_i + 1}{\hat{v}_i} + \sin y_i \frac{\hat{a} \sin x_i}{\hat{v}_i}$$

where $A_1(\kappa) = I_1(\kappa)/I_0(\kappa)$. These solutions must be computed numerically and, in our experience, Newton-Raphson type algorithms can occasionally be unstable.

Standard errors for $\hat{a}$ and $\hat{b}$ can be obtained from the Hessian matrix in the usual way.

4. Generalizations

Note that, unlike the Sine and Cosine models, our regression model is not symmetric in x and y . In modelling toroidal data, this may give the opportunity to select the explanatory and response variables. In this section we propose a *polynomial*-like generalization, and a multiple regression formulation. Unfortunately, in both cases the maximum likelihood estimates must be computed by direct function evaluation.

Given $(x_1, y_1), \dots, (x_n, y_n)$ then a polynomial model of order k is given by

$$Y | x \sim \text{vM} \left(\text{atan2} \left(a_1 \sin x + \dots + a_k \sin^k x, B \cos x + k \right), \kappa \right). \quad (2)$$

where $a_1, \dots, a_k, B$ and κ are parameters to be estimated.

For multiple regression the model is

$$Y | \mathbf{x} \sim \text{vM} \left(\text{atan2} \left(\mathbf{a}^T \sin \mathbf{x}, \mathbf{b}^T \cos \mathbf{x} + p \right), \kappa \right) \quad (3)$$

where $\mathbf{x}^T = (x_1, \dots, x_p)$, and $\kappa, \mathbf{a}^T = (a_1, \dots, a_p)$ and $\mathbf{b}^T = (b_1, \dots, b_p)$ are parameters to be estimated.

5. Example

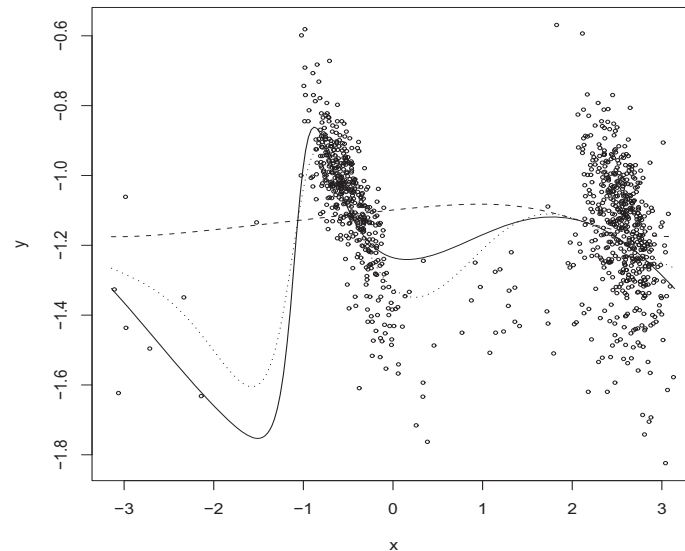
The *backbone* of a protein comprises a sequence of atoms, $\text{N}_1 - \text{C}_1^\alpha - \text{C}_1 - \dots - \text{N}_m - \text{C}_m^\alpha - \text{C}_m$, in which each group $\text{N}_i - \text{C}_i^\alpha - \text{C}_i$ is associated with a pair of *dihedral* angles and a type

of amino acid. We illustrate the above methods for pairs associated with the amino acid *proline*. A scatter plot of 1000 observations is given in Figure 2, to which we have fit polynomial models of order 1, 2, and 3. The parameter estimates, with the standard errors are given in Table 1, and the “quadratic” model is selected by a forward stepwise rule.

Table 1: *Parameter estimates (and SEs) for proline data using polynomial model (2)*

$\hat{B}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{\kappa}$	-LL
-0.447 (.156)	-0.042 (.009)			31.21 (1.384)	983.8
1.910 (.011)	-0.267 (.005)	-0.455 (.010)		44.77 (1.991)	988.7
2.642 (.034)	-0.570 (.011)	-0.608 (.020)	0.556 (.030)	46.30 (2.059)	989.0

Figure 2: *Regression lines from model (2) with $k = 1$ (dashed), $k = 2$ (continuous) and $k = 3$ (dotted).*



References

- Downs T.D. and Mardia K.V. (2002) Circular regression, *Biometrika*, 89, 683–397.
- Hughes G. (2007) *Multivariate and Time Series Models for Circular Data with Applications to Protein Conformational Angles*, Ph.D. thesis, Department of Statistics, University of Leeds, Leeds, U.K.
- Jammalamadaka S.R. and SenGupta A. (1998) Predictive inference for directional data, *Statistics & Probability Letters*, 40, 247–257.
- Kato S., Shimizu K. and Shieh G.S. (2008) A circular-circular regression model, *Statistica Sinica*, 18, 633–64.
- Kent J.K. and Mardia K.V. (2002) (discussion of) spatial-temporal nonlinear filtering based on hierarchical statistical models, *Test*, 11, 287–288.
- Mardia K.V. and Jupp P.E. (2000) *Directional Statistics*, Wiley, Chichester, U.K.