Indicators for effects on mean and variance in projected normal regression models for a circular outcome

ARTICLE HISTORY
Compiled September 26, 2018

ABSTRACT
Projected normal models for circular data are so called heterogeneous error models, the structure of the model allows for the simultaneous modelling of effects on the mean (a location effect) and variance (an accuracy effect) of a circular variable. In this paper we investigate several measures for assessing whether there is either just an effect on the variance or both an effect on the variance and the mean of the circular outcome in a circular regression model. In previous literature a measure to do so, the signed shortest distance to the origin ($SSDO$), has already been introduced. However, our simulations suggest that the type I error of the $SSDO$ is greater than the specified α level. We introduce a hypothesis test based on an angle based measure that achieves, in our simulations, the correct type I error.

KEYWORDS
circular regression; projected normal models; heterogeneous error

1. Introduction

Circular regression models are those in which a circular variable $\theta \in [0, 2\pi)$ is regressed on a set of linear and/or circular predictors, $\boldsymbol{x}$. In the literature there are three approaches to circular data, the 'intrinsic', 'wrapping' and 'embedding' approach [1]. In the intrinsic approach models are based on distributions directly defined on the circle whereas in the wrapping and embedding approach distributions are defined in respectively $\mathbb{R}$ and $\mathbb{R}^2$ and subsequently wrapped or projected onto the circle. Within each of these three approaches regression models for a circular outcome have been introduced [2–7].

Another distinction that can be made in the literature on circular regression models is one between models with homogeneous errors and those with heterogeneous errors [8]. In a homogeneous error model predictor variables can only affect the mean of the circular outcome while in heterogeneous error models predictor variables can have an effect on both the mean (a location effect) and variance (an accuracy effect) of the circular outcome. The projected normal (PN) regression model, the model we focus on in this paper, is one of these heterogeneous error models for circular variables. This model was first introduced by [4] and adapted to the Bayesian context by [7].

Several measures for assessing the effect on the mean of the outcome, a location effect, in a PN regression model were introduced in [9]. Additionally they introduced an indicator for effects on both the variance and the mean. This indicator is called the $SSDO$. In this paper new indicators for effects on the variance and the mean in PN regression models will be introduced in Section 3. As for the $SSDO$ these new indicators will test H_0 : *There is only an accuracy effect.* The performance of the new and existing indicators will be assessed in a simulation study in Section 4. We will

however first give a short introduction to the PN normal regression model in Section 2.

2. Projected Normal Regression Models for a Circular Outcome

In projected normal models we assume that the circular outcome θ results from a projection onto the circle of a bivariate normal variable $\boldsymbol{y}_i \sim N_2(\boldsymbol{\mu}_i, \boldsymbol{I})$ where $i, \ldots, n$. The relation between $\boldsymbol{y}$ and θ is defined as:

$$\boldsymbol{u} = \boldsymbol{y}/r \tag{1}$$

where $\boldsymbol{u} = (\cos\theta, \sin\theta)^t$ and $r > 0$. The idea behind this projection is that we do not have to conduct inference on θ directly but we can indirectly conduct inference on a bivariate normal variable $\boldsymbol{y}$. This makes for a flexible approach as a lot of different and complex models exist for bivariate normal data. However, both $\boldsymbol{y}$ and r cannot be directly obtained from θ. Instead the estimation of $\boldsymbol{y}$ and r is treated as a missing data problem. In this paper we use the approach used by [9] to solve the missing data problem and fit the model.

The relation in (1) implies that θ has a projected normal distribution defined as:

$$PN(\theta \mid \boldsymbol{\mu}, \boldsymbol{I}) = \frac{1}{2\pi} e^{-\frac{1}{2}||\boldsymbol{\mu}||^2} \left[1 + \frac{\boldsymbol{u}^t\boldsymbol{\mu}\Phi(\boldsymbol{u}^t\boldsymbol{\mu})}{\phi(\boldsymbol{u}^t\boldsymbol{\mu})}\right], \tag{2}$$

where $-\pi \leq \theta < \pi$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^t \in \mathbb{R}$ is the mean vector, the covariance matrix $\boldsymbol{I}$ is identity and $\boldsymbol{u}^t = (\cos\theta, \sin\theta)$. The terms $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution. Note that we choose the covariance matrix to be identity for identification purposes. In Figure 1 we see that the shape of this density is rotationally symmetric about its mean direction $\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$ and its concentration is dependent on $||\ \boldsymbol{\mu}\ ||^2$ (see [10] for the exact form of this relation). A different way of parameterizing a projected normal distribution can be found in [11].

In a circular regression model $\boldsymbol{\mu}$ may have the following structure:

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_i^I \\ \mu_i^{II} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{\beta}^I)^t\boldsymbol{x}_i^I \\ (\boldsymbol{\beta}^{II})^t\boldsymbol{x}_i^{II} \end{pmatrix}, \tag{3}$$

where $i = 1, \ldots, n$, $\boldsymbol{x}_i$ is a vector of predictor values for individual i and each $\boldsymbol{\beta}$ is a vector with intercept and regression coefficients. To be able to estimate an intercept, the first component of $\boldsymbol{x}_i$ equals 1. In this paper we center the predictor x and estimate the PN regression model using MCMC methods also used in [9]. Because the structure in (3) is such that the predictor variables determine $\boldsymbol{\mu}$ the predictors can have effects on both the mean (a location effect) and the variance (an accuracy effect) of the circular outcome.

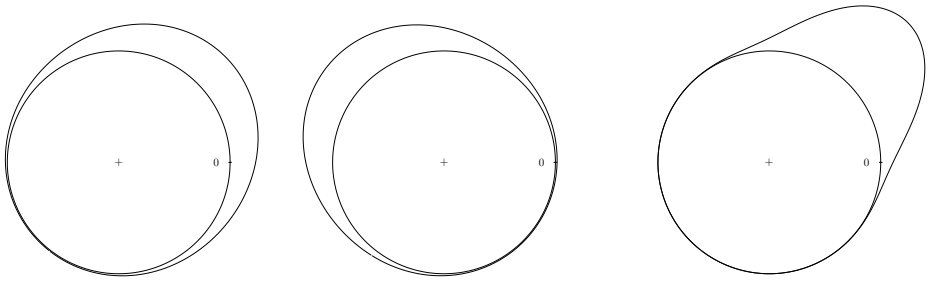

Figure 1. Projected normal densities with different $\boldsymbol{\mu}$. From left to right, the mean vector $\boldsymbol{\mu}$ is set to (2,2), (-2,2) and (4,4).

3. Two indicators for accuracy and location effects

In this section we will describe two different indicators that allow us to check whether the predictors in a projected normal regression model only have an effect on the variance (an accuracy effect) or also have an effect on the location (a location effect) of the circular outcome. The first indicator, the $SSDO$, has been introduced previously in [9]. The second indicator is new and actually comprises a set of indicators that we call angle based measures.

3.1. The signed shortest distance to the origin (SSDO)

In [9] it is outlined that the effects of a variable x on both the variance and mean of the circular outcome in a PN regression model can be detected by looking at the shortest distance of the regression line in bivariate space to the origin (SDO). The regression line in $\mathbb{R}^2$ is defined as follows:

$$(\beta_0^I + \beta_1^I x,\ \beta_0^{II} + \beta_1^{II} x)^t.$$

Figure 2 shows two regression lines in $\mathbb{R}^2$ together with arrows representing predicted outcomes for $x = (x_{min}, 0, x_{max})$ and a dotted line that represents the SDO. The intersections of the arrows with the circle represent the predicted outcomes on the circle. The left figure shows a situation with only an accuracy effect. In this situation the regression line runs through the origin ($SDO = 0$) and the arrows for the predicted outcomes on the circle are parallel and they intersect with the circle at one point, i.e.: the circular predicted values for different values of x are the same. The right figure shows a situation where there is also a location effect. In this situation the regression line does not run through the origin, $SDO > 0$ and the arrows representing the predicted outcomes on the circle are not parallel and do not intersect the circle at the same point, i.e.: the circular predicted values for different values of x are different.

The authors in [9] introduced a measure, the signed shortest distance to the origin ($SSDO$), derived from the SDO, to detect accuracy and location effects. Just as for

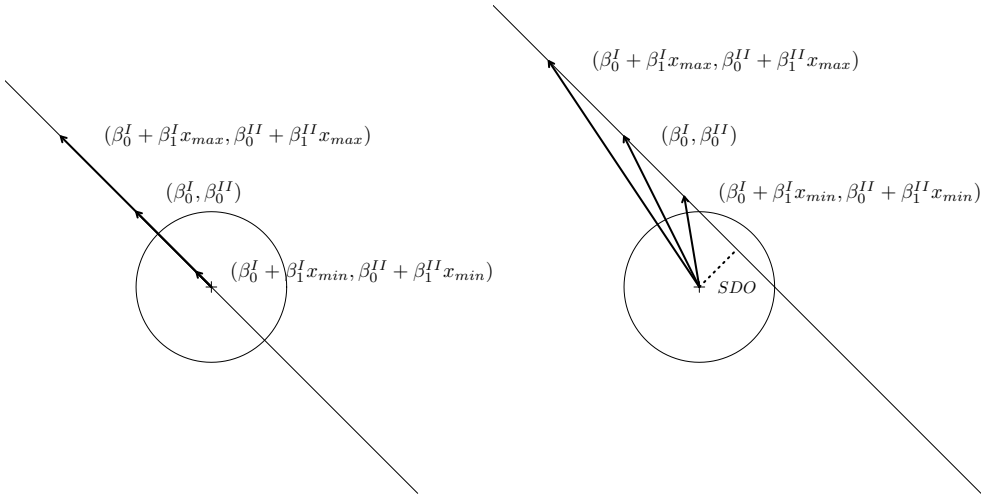

Figure 2. Two regression lines in $\mathbb{R}^2$ together with a unit circle and arrows representing predicted outcomes for $x = (x_{min}, 0, x_{max})$ and the SDO.

the SDO, H_0 : *There is only an accuracy effect* is true for the $SSDO$ when it equals 0. To test whether there is also a location effect with the $SSDO$ we can thus reformulate this null hypothesis as $H_0 : SSDO = 0$. The authors in [9] make use of the highest posterior density (HPD) to test the null hypothesis. The null hypothesis is rejected at a significance level of α if the $100 * (1 - \alpha)$ HPD interval for SSDO does not include zero.

The authors in [9] perform a simulation for the $SSDO$. This simulation is however limited to two sample sizes and does not adequately investigate the validity of the test. With validity we mean achieving a correct type-I error in the long run. We will therefore perform a new simulation with more sample sizes in Section 4. Additionally, we will set up this simulation such that it contains a larger variation of real SDO values than before. Finally, we also include the new angle based measures (introduced next) in this simulation. This new set-up will allow us to assess and compare the type-I errors of the different measures.

3.2. Angle based measures

In Figure 2 we saw that if the vectors representing predicted values for several x are not parallel there is a location effect. The new measures we introduce in this section are based on testing whether these vectors are parallel. A way to test whether two vectors are parallel is to compute the sine of the angle between the two vectors and check whether this is equal to 0. The sine of the angles λ and γ between the vectors $(\beta_0^I, \beta_0^{II})$ and $(\beta_0^I + \beta_1^I x_{min}, \beta_0^I + \beta_1^{II} x_{min})$ and $(\beta_0^I + \beta_1^I x_{max}, \beta_0^I + \beta_1^{II} x_{max})$ respectively from Figure 2 are computed as follows:

$$\sin(\lambda) = \frac{\beta_0^I(\beta_0^{II} + \beta_1^{II} x_{min}) - \beta_0^{II}(\beta_0^I + \beta_1^I x_{min})}{\sqrt{(\beta_0^I)^2 + (\beta_0^{II})^2}\sqrt{(\beta_0^I + \beta_1^I x_{min})^2 + (\beta_0^{II} + \beta_1^{II} x_{min})^2}} \tag{4}$$

$$\sin(\gamma) = \frac{\beta_0^I(\beta_0^{II} + \beta_1^{II}x_{max}) - \beta_0^{II}(\beta_0^I + \beta_1^I x_{max})}{\sqrt{(\beta_0^I)^2 + (\beta_0^{II})^2}\sqrt{(\beta_0^I + \beta_1^I x_{max})^2 + (\beta_0^{II} + \beta_1^{II}x_{max})^2}} \tag{5}$$

and are bounded between -1 and 1. From now on we will call these measures $\sin(\lambda)$ and $\sin(\gamma)$. Note that in this paper, because the predictor x has been centered, the vector $(\beta_0^I, \beta_0^{II})$ is the vector pointing at the data mean. We can thus reformulate H_0 : *There is only an accuracy effect* as $H_0 : \sin(\lambda) = 0$ or $H_0 : \sin(\gamma) = 0$. We can test this null hypothesis by using the 95% HPD interval of the posterior distribution of $\sin(\lambda)$ or $\sin(\gamma)$. We could also measure the angle between the vectors for the predicted values at x_{min} and x_{max}. A reason to compute this other measure is that the angle is always larger than either λ or γ on its own, which means that it may also have larger power for rejecting H_0. We define this measure as follows:

$$\sin(\lambda{+}\gamma) = \frac{(\beta_0^I + \beta_1^I x_{min})(\beta_0^{II} + \beta_1^{II}x_{max}) - (\beta_0^{II} + \beta_1^{II}x_{min})(\beta_0^I + \beta_1^I x_{max})}{\sqrt{(\beta_0^I + \beta_1^I x_{min})^2 + (\beta_0^{II} + \beta_1^{II}x_{min})^2}\sqrt{(\beta_0^I + \beta_1^I x_{max})^2 + (\beta_0^{II} + \beta_1^{II}x_{max})^2}} \tag{6}$$

The three measures we have just introduced, $\sin(\lambda)$, $\sin(\gamma)$ and $\sin(\lambda + \gamma)$, are all dependent on the shape of the data. For testing purposes we would rather consider a measure that is not dependent on the shape of the data. An intuitive way to construct such a measure is to consider the vector $(\beta_1^I, \beta_1^{II})$ and test whether this is parallel to $(\beta_0^I, \beta_0^{II})$:

$$\sin\beta = \frac{\beta_0^I\beta_1^{II} - \beta_0^{II}\beta_1^I}{\sqrt{(\beta_0^I)^2 + (\beta_0^{II})^2}\sqrt{(\beta_1^I)^2 + (\beta_1^{II})^2}}. \tag{7}$$

From now on we will call this measure $\sin(\beta)$. A visual representation of β is given in the left plot in Figure 3. The right plot shows how $(\beta_1^I, \beta_1^{II})$ is related to the regression line in bivariate space; the addition of $(\beta_1^I, \beta_1^{II})$ to $(\beta_0^I, \beta_0^{II})$ results in the vector $(\beta_0^I + \beta_1^I, \beta_0^I + \beta_1^{II})$.

Note that for all measures introduced above we could also use the cosine formula, $\cos(\theta) = \boldsymbol{a} \cdot \boldsymbol{b}/||\boldsymbol{a}||||\boldsymbol{b}||)$, to compute the angle between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. However, the cosine of the angle between two vectors is -1 or 1 if they are parallel, which is not convenient for the purpose of testing purposes.

4. Simulation Study

To assess the performance of $SSDO$ and the angle based measures $\sin(\beta)$, $\sin(\gamma)$, $\sin(\lambda)$ and $\sin(\lambda+\gamma)$ we conducted a simulation study with 477 designs. In 72 of these designs, the accuracy designs, there was only an accuracy effect ($SDO = 0$). In the other designs, the location designs, there was also a location effect $SDO > 0$.

For each design 2500 datasets were simulated. These datasets had different sample sizes (N): 25, 50, 75, 100 and 150. These sample sizes were chosen because in earlier simulations for the $SSDO$ in [9] only two sizes (50 and 200) were used and in this

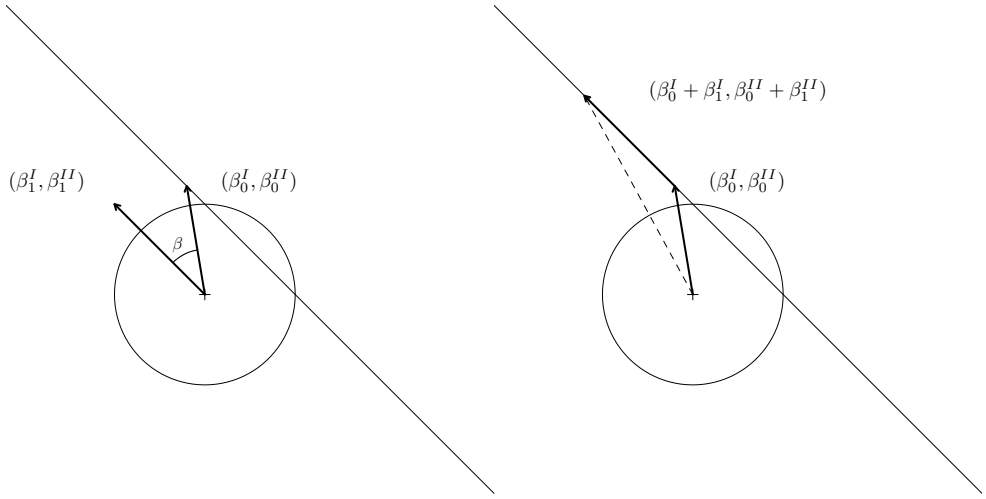

Figure 3. Visual representation of $\sin(\beta)$.

paper we want to see more detailed results regarding sample size. For each sample size 500 datasets were simulated. Each dataset contains one circular outcome θ and one linear predictor $x \sim N(0, 1)$ similar to the simulation in [9]. The relation between predictor and outcome was determined by the population values for the intercepts, β_0^I and β_0^{II}, and coefficients, β_1^I and β_1^{II}. In each design different population values were chosen for the intercepts and regression coefficients. For the designs with a location effect pairs of population values $(\beta_0^I, \beta_0^{II})$ for the linear intercepts were:

$$\{(\cos(10°), \sin(10°)), (\cos(30°), \sin(30°)), (\cos(45°), \sin(45°)), (\cos(60°), \sin(60°)), (\cos(80°), \sin(80°))\}$$

and pairs of population values for the regression coefficients $(\beta_1^I, \beta_1^{II})$ were:

$$\{(1, 0), (0, 1)(1, 1), (0.5, 0), (0, 0.5), (0.5, 0.5), (2, 0), (0, 2), (2, 2)\}.$$

These population values were then multiplied by a multiplication factor between 1 and 5 at intervals of 0.5 to obtain regression lines with different SDO. Note that the coefficients were chosen to be largely similar to the coefficients that show good performance (in terms of bias and coverage) in previous simulation studies in [9] and [12]. Datasets for each possible combination of the pairs of intercepts and regression coefficients and multiplication factors were simulated. These combinations led to a larger variation in real SDO values than in previous simulations. In the accuracy designs the population values for β_0^I and β_1^I and β_0^{II} and β_1^{II} were equal in each design. Their exact values and the simulation code is given in the supplementary material.

From the population values of the intercepts and coefficients, the population values of the SDO, $SSDO$ and the angle based measures ($\sin(\beta)$, $\sin(\lambda)$ and $\sin(\gamma)$) were computed. For each dataset we determine whether x is predicted to have an accuracy effect by checking whether the 95% HPD intervals of the estimated $SSDO$ and the angle based measures included 0. We thus test H_0 : *There is only an accuracy effect.* If H_0 was not rejected the dataset was classified as having only an accuracy effect. For each design we then computed the proportion of datasets in which the $SSDO$ and the

Table 1. The proportion of datasets in which an accuracy effect is indicated by $SSDO$ grouped per real SDO value.

N	$SDO = 0$	$0 < SDO < 1$	$1 \leq SDO < 2$	$2 \leq SDO < 3$	$3 \leq SDO < 4$	$4 \leq SDO$
25	0.90	0.49	0.23	0.14	0.10	0.06
50	0.89	0.35	0.15	0.10	0.08	0.05
75	0.90	0.28	0.12	0.08	0.07	0.05
100	0.90	0.25	0.10	0.07	0.06	0.04
150	0.89	0.19	0.08	0.06	0.06	0.04

angle based measures indicated only an accuracy effect.

To display the results of the simulation in a concise manner, we grouped all simulation designs into 6 categories based on their population SDO: those where $SDO = 0$ (the 72 accuracy designs) and those where the $0 < SDO < 1$ (72 designs), $1 \leq SDO < 2$ (90 designs), $2 \leq SDO < 3$ (90 designs), $3 \leq SDO < 4$ (90 designs) and $4 \leq SDO$ (63 designs). For each of the 6 categories of our simulation we then averaged the proportion of datasets where only an accuracy effect was indicated over the designs of that category. For the location designs this proportion represents a type-II error; the proportion in which H_0 : *There is only an accuracy effect.* was not rejected even though it should have been. For the accuracy designs this proportion is equal to 1 minus the type-I error of the test or actually the proportion in which we correctly classify the effect as having only an accuracy effect. It also represents the coverage of the HPD interval; we are testing whether the interval includes 0 which is the real value in the accuracy designs. Thus, if the $SSDO$ and angle based measures perform well we expect the proportion to be high, around 0.95 because we set $\alpha = 0.05$, for the accuracy designs and low for the designs with a location effect.

4.1. Results

The results of the simulation are shown in Tables 1 and 2 for the performance of the $SSDO$ and $\sin(\beta)$ respectively. The tables show the proportion of datasets in which H_0 : *There is only an accuracy effect* could not be rejected grouped per real SDO value.

We see that for the location designs ($SDO > 0$) both measures perform better with increasing sample sizes and with increasing SDO. The proportion of datasets in which H_0 is not rejected ($SSDO = 0$ or $\sin(\beta) = 0$) becomes lower when sample size and SDO increase. For the accuracy designs there is no clear effect of sample size for the $SSDO$. The proportion of datasets in which H_0 is not rejected stays the same at around 0.9. It thus does not reach the correct type-I error level (of 0.05) within the current range of sample sizes. The fact that the actual alpha level consistently seems to lie at 0.1 is not promising of the tests ability to reach the correct alpha level asymptotically. There is an effect of sample size for $\sin(\beta)$ in the accuracy designs. This indicator reaches the correct type-I error level at a sample size of 75 meaning that in the long run this is a valid test for H_0.

Results for the other three angle based measures, $\sin(\gamma)$, $\sin(\lambda)$ and $\sin(\lambda + \gamma)$, are shown in the Appendix. In general these measures perform worse or equal to $\sin(\beta)$. Only the performance of $\sin(\gamma)$ stands out. This measure already reaches the correct type-I error level for a sample size of 25.

Table 2. The proportion of datasets in which an accuracy effect is indicated by $\sin(\beta)$ grouped per real SDO value.

N	$SDO = 0$	$0 < SDO < 1$	$1 \leq SDO < 2$	$2 \leq SDO < 3$	$3 \leq SDO < 4$	$4 \leq SDO$
25	0.91	0.52	0.27	0.17	0.13	0.08
50	0.92	0.39	0.18	0.12	0.09	0.06
75	0.94	0.33	0.14	0.09	0.07	0.05
100	0.95	0.29	0.12	0.08	0.07	0.05
150	0.94	0.23	0.10	0.07	0.06	0.04

5. Discussion

In this paper we have introduced new angle based measures that allow testing of H_0 : *There is only an accuracy effect* in projected normal regression models for a circular outcome. We have investigated their performance and compared these measures to a previously introduced measure, the $SSDO$ in a simulation study.

The results of the simulation study lead us to conclude that in general the angle based measures perform better than the $SSDO$ since three of them reach the correct type-I error level in the long run. However, three of the four new measures that we introduced in this paper, $\sin(\gamma)$, $\sin(\lambda)$ and $\sin(\lambda+\gamma)$, are dependent on the distribution of the predictor variable. In further research we could focus on finding more angle based measures that are not dependent on the distribution of the predictor variable. To do so standardization of the predictor x instead of centering only could be an option. Additionally now that we have found tests that have valid type-I errors in the long run we can try to increase the power of these tests.

Lastly, we would like to note that in this paper we have only considered simulation designs in which there is an effect in the population, either only an accuracy effect or both an accuracy and a location effect. In reality however there are situations in which there is no effect at all. In further research it would be beneficial to investigate the performance of the $SSDO$ and the new measures introduced here in samples for which there is no effect in the population.

6. References

References

[1] Mardia KV, Jupp PE. Directional statistics. Chichester, England: Wiley; 2000.

[2] Fisher NI, Lee AJ. Regression models for an angular response. Biometrics. 1992;48(3):665–677.

[3] Lagona F. Regression analysis of correlated circular data based on the multivariate von mises distribution. Environmental and ecological statistics. 2016;23(1):89–113.

[4] Presnell B, Morrison SP, Littell RC. Projected multivariate linear models for directional data. Journal of the American Statistical Association. 1998;93(443):1068–1077.

[5] Mulder K, Klugkist I. Bayesian estimation and hypothesis tests for a circular generalized linear model. Journal of Mathematical Psychology. 2017;80:4–14.

[6] Ravindran P, Ghosh SK. Bayesian analysis of circular data using wrapped distributions. Journal of Statistical Theory and Practice. 2011;5(4):547–561.

[7] Nuñez-Antonio G, Gutiérrez-Peña E, Escarela G. A Bayesian regression model for circular data based on the projected normal distribution. Statistical Modelling. 2011;11(3):185–201.

[8] Rivest LP, Duchesne T, Nicosia A, et al. A general angular regression model for the analysis of data on animal movement in ecology. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2015;63(3):445–463.
[9] Cremers J, Mulder K, Klugkist I. Circular interpretation of regression coefficients. British Journal of Mathematical and Statistical Psychology. 2018;71(1):75–95.
[10] Kendall DG. Pole-seeking brownian motion and bird navigation. Journal of the Royal Statistical Society Series B. 1974;37:97–133.
[11] Wang F, Gelfand AE. Directional data analysis under the general projected normal distribution. Statistical Methodology. 2013;10(1):113–127.
[12] Cremers J, Mainhard MT, Klugkist I. Assessing a bayesian embedding approach to circular regression models. Methodology. 2018;14(2):69–81.

7. Appendices

Table 1. The proportion of datasets in which an accuracy effect is indicated by $\sin(\lambda)$ grouped per real SDO value.

N	$SDO = 0$	$0 < SDO < 1$	$1 \leq SDO < 2$	$2 \leq SDO < 3$	$3 \leq SDO < 4$	$4 \leq SDO$
25	0.90	0.50	0.25	0.16	0.12	0.07
50	0.89	0.38	0.16	0.11	0.08	0.05
75	0.92	0.31	0.13	0.09	0.07	0.05
100	0.93	0.27	0.11	0.08	0.07	0.05
150	0.93	0.22	0.09	0.07	0.06	0.04

Table 2. The proportion of datasets in which an accuracy effect is indicated by $\sin(\gamma)$ grouped per real SDO value.

N	$SDO = 0$	$0 < SDO < 1$	$1 \leq SDO < 2$	$2 \leq SDO < 3$	$3 \leq SDO < 4$	$4 \leq SDO$
25	0.94	0.56	0.28	0.18	0.13	0.08
50	0.94	0.42	0.18	0.12	0.09	0.05
75	0.95	0.34	0.14	0.09	0.07	0.05
100	0.95	0.30	0.12	0.08	0.07	0.05
150	0.95	0.23	0.10	0.07	0.06	0.04

Table 3. The proportion of datasets in which an accuracy effect is indicated by $\sin(\gamma + \lambda)$ grouped per real SDO value.

N	$SDO = 0$	$0 < SDO < 1$	$1 \leq SDO < 2$	$2 \leq SDO < 3$	$3 \leq SDO < 4$	$4 \leq SDO$
25	0.91	0.53	0.25	0.16	0.12	0.07
50	0.92	0.39	0.17	0.11	0.08	0.05
75	0.94	0.32	0.13	0.09	0.07	0.05
100	0.95	0.28	0.11	0.08	0.07	0.05
150	0.95	0.23	0.09	0.07	0.06	0.04