CircNNTSR: An **R** Package for the Statistical Analysis of Circular, Multivariate Circular, and Spherical Data Using Nonnegative Trigonometric Sums

Juan José Fernández-Durán
Instituto Tecnológico Autónomo de México

María Mercedes Gregorio-Domínguez
Instituto Tecnológico Autónomo de México

Abstract

The statistical analysis of circular, multivariate circular, and spherical data is very important in different areas, such as paleomagnetism, astronomy and biology. The use of nonnegative trigonometric sums allows for the construction of flexible probability models for these types of data to model datasets with skewness and multiple modes. The R package **CircNNTSR** includes functions to plot, fit by maximum likelihood, and simulate models based on nonnegative trigonometric sums for circular, multivariate circular, and spherical data. For maximum likelihood estimation of the models for the three different types of data an efficient Newton-like algorithm on a hypersphere is used. Examples of applications of the functions provided in the **CircNNTSR** package to actual and simulated datasets are presented and it is shown how the package can be used to test for uniformity, homogeneity, and independence using likelihood ratio tests.

Keywords: Fourier series, likelihood ratio test, maximum likelihood estimation, smooth Riemann manifold.

1. Introduction

There exist random variables that have supports that differ from a subset of the real line. For example, the direction taken by an animal and the wind direction correspond to random variables with support corresponding to the circumference of the unit circle. In these cases,

the random variable θ is an angle with possible values on the circumference of the unit circle in the interval $(0, 2\pi]$. Therefore, the density function of a circular random variable must be periodic. By considering a vector of circular random variables, $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_R)$, one can analyze multivariate circular data. Examples of multivariate circular data include the pair of dihedral angles in the backbone of a protein and the wind directions at different monitoring stations. Spherical data consist of a bivariate vector $\underline{\theta}_s = (\theta_{s1}, \theta_{s2})$ with $\theta_{s1} \in (0, 2\pi]$ and $\theta_{s2} \in (0, \pi]$. In the case of data for the Earth's surface, θ_{s1} is identified with a longitudinal coordinate and θ_{s2} is identified with a latitudinal coordinate. Examples of spherical data include the position of the occurrence of different events on the Earth's surface such as the epicenter of an earthquake or the trajectory of a hurricane. The time at which an event occurs, such as the epileptic seizures of a patient during the day, can also be analyzed as circular data. Regarding possible applications, Ridout and Linkie (2009), Linkie and Ridout (2011), and Lynam *et al.* (2013), for example, used circular models to study the activity patterns of wild cats using camera-trap data.

Fernández-Durán (2004) developed a family of univariate densities for circular data based on nonnegative trigonometric sums (Fejér 1916). This family is referred to as NNTS models. In terms of complex numbers, the probability density function of an NNTS model can be expressed as follows:

$$f(\theta) = \left\| \sum_{k=0}^{M} c_k e^{ik\theta} \right\|^2 = \sum_{k=0}^{M} \sum_{m=0}^{M} c_k \bar{c}_m e^{i(k-m)\theta}, \tag{1}$$

where $i = \sqrt{-1}$, $e^{ik\theta} = \cos(k\theta) + i\sin(k\theta)$, $c_k = c_{rk} + ic_{ik}$ is a complex number with c_{rk} and c_{ik} being the real and imaginary components of c_k, respectively, and $\bar{c}_k = c_{rk} - ic_{ik}$ represents the conjugate of c_k. Then, $f(\theta)$ is the squared norm of a sum of complex numbers and is thus nonnegative. In order for $f(\theta)$ to integrate to one, the following restriction on the c parameters must be imposed:

$$\sum_{k=0}^{M} \|c_k\|^2 = \frac{1}{2\pi}, \tag{2}$$

implying that $c_{i0} = 0$, i.e., c_0 must be a real number. Given this restriction on the c parameters, the case of $M = 0$ corresponds to the uniform circular density, $f(\theta) = \frac{1}{2\pi}$. The number of free c parameters is equal to $2M$.

This idea of using the squared norm of a sum of complex numbers to define a density function can be extended to specify the joint density of a vector of circular random variables in the case of multivariate circular data and the bivariate random vector that defines spherical data. The NNTS models for multivariate circular and spherical data are referred to as MNNTS and SNNTS models, respectively. In the case of multivariate circular data, the squared norm of a multiple complex trigonometric sum is used. For a bivariate circular random vector, $\underline{\theta} = (\theta_1, \theta_2)$, let

$$
\begin{aligned}
f(\underline{\theta}) = f(\theta_1, \theta_2) &= \left\| \sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} c_{k_1 k_2} e^{i(k_1 \theta_1 + k_2 \theta_2)} \right\|^2 \\
&= \sum_{k_1=0}^{M_1} \sum_{m_1=0}^{M_1} \sum_{k_2=0}^{M_2} \sum_{m_2=0}^{M_2} c_{k_1 k_2} \bar{c}_{m_1 m_2} e^{i[(k_1 - m_1)\theta_1 + (k_2 - m_2)\theta_2]}.
\end{aligned}
\tag{3}
$$

In order for $f(\theta_1, \theta_2)$ to integrate to one, the following restriction on the c parameters must be imposed:

$$\sum_{k=0}^{M_1} \sum_{m=0}^{M_2} \|c_{km}\|^2 = \left(\frac{1}{2\pi}\right)^2. \tag{4}$$

The support of a bivariate circular random vector is the surface of a torus. This result can be extended to circular random vectors of dimension R greater than two by defining

$$
\begin{aligned}
f(\underline{\theta}) = f(\theta_1, \ldots, \theta_R) &= \left\| \sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} \cdots \sum_{k_R=0}^{M_R} c_{k_1 \cdots k_R} e^{i(k_1\theta_1 + \cdots + k_R\theta_R)} \right\|^2 \\
&= \sum_{k_1=0}^{M_1} \sum_{m_1=0}^{M_1} \cdots \sum_{k_R=0}^{M_R} \sum_{m_R=0}^{M_R} c_{k_1 \cdots k_R} \bar{c}_{m_1 \cdots m_R} e^{i[(k_1-m_1)\theta_1 + \cdots + (k_R-m_R)\theta_R]},
\end{aligned}
\tag{5}
$$

with the following restriction on the c parameters:

$$\sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} \cdots \sum_{k_R=0}^{M_R} \|c_{k_1 \cdots k_R}\|^2 = \left(\frac{1}{2\pi}\right)^R. \tag{6}$$

Given this restriction, the parameter space of NNTS and MNNTS densities correspond to the surface of a hypersphere. Fernández-Durán and Gregorio-Domínguez (2014b) demonstrated that the marginal and conditional densities of an MNNTS joint density are members of the MNNTS family. The number of free c parameters of an MNNTS model is equal to $2(\prod_{k=1}^{R}(M_k + 1)) - 2$.

In the case of spherical data, $\underline{\theta}_s = (\theta_{s1}, \theta_{s2})$ with $\theta_{s1} \in (0, 2\pi]$ and $\theta_{s2} \in (0, \pi]$. The SNNTS density is defined as follows:

$$
\begin{aligned}
f(\underline{\theta}_s) &= \sin(\theta_{s2}) \left\| \sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} c_{k_1 k_2} e^{i(k_1\theta_{s1} + k_2\theta_{s2})} \right\|^2 \\
&= \sin(\theta_{s2}) \sum_{k_1=0}^{M_1} \sum_{m_1=0}^{M_1} \sum_{k_2=0}^{M_2} \sum_{m_2=0}^{M_2} c_{k_1 k_2} \bar{c}_{m_1 m_2} e^{i[(k_1-m_1)\theta_{s1} + (k_2-m_2)\theta_{s2}]}.
\end{aligned}
\tag{7}
$$

The term $\sin(\theta_{s2})$ corresponds to the uniform measure on the sphere. The restriction in the parameter space corresponds to the following:

$$2\pi \sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} \sum_{\substack{m_2=0 \\ |k_2-m_2|\neq 1}}^{M_2} c_{k_1 k_2} \bar{c}_{k_1 m_2} \left(\frac{1 + \cos((k_2-m_2)\pi)}{1 - (k_2-m_2)^2} \right) = 1. \tag{8}$$

The restriction in Equation 8 can also be written as

$$\sum_{k_1=0}^{M_1} \sum_{k_2=0}^{M_2} \|c_{k_1 k_2}^*\| = 1, \tag{9}$$

by a transformation of the parameter space (for details see Fernández-Durán and Gregorio-Domínguez 2014a). Therefore, the new transformed space corresponds again to a hypersphere,

and the NNTS, MNNTS and SNNTS cases can be developed in a similar manner. The number of free c parameters in an SNNTS model is equal to $2(M_1 + 1)(M_2 + 1) - 2$.

2. Maximum likelihood estimation and likelihood ratio tests

Maximum likelihood estimation of the c parameters of the NNTS, MNNTS, and SNNTS models is implemented numerically by using a Newton-like algorithm on manifolds. In the NNTS, MNNTS, and transformed SNNTS cases, the c parameter space (smooth Riemann manifold) corresponds to a hypersphere as defined by the restrictions in Equations 2, 4, 6, and 9. This algorithm is described in detail in Fernández-Durán and Gregorio-Domínguez (2010b) for circular data. The modifications for multivariate and spherical data are directly obtainable and can be consulted in Fernández-Durán and Gregorio-Domínguez (2014a,b).

By using the maximum likelihood method of estimation, it is possible to apply different likelihood ratio tests. In practice, tests of uniformity, homogeneity, and independence are among the most important tests for circular, multivariate circular, and spherical data. The uniformity null hypothesis states that the data follow a uniform distribution. Given the definition of NNTS, MNNTS, and SNNTS models, it is equivalent to test that the M parameters are equal to zero. For example, in the case of circular data, the uniformity null hypothesis is equivalent to $H_0 : M = 0$ and a likelihood ratio test can be implemented by comparing the maximized log-likelihood of the uniform model, which is an NNTS model with $M = 0$, against the maximized log-likelihood of an NNTS model with $M > 0$. The likelihood-ratio test statistic Λ_{Unif} (see Fernández-Durán and Gregorio-Domínguez 2014c) asymptotically follows a chi-squared distribution with twice as many degrees of freedom as the number of free c parameters in the non-uniform model. The same definition can be used for multivariate circular and spherical uniform tests. One disadvantage of the NNTS test of uniformity is the need to select a value for M which is related to the maximum number of peaks considered to exist in the population. Fernández-Durán and Gregorio-Domínguez (2014c) compare the NNTS uniformity test with other more conventional tests and recommend to select the larger value when considering two consecutive values for M. In the cases of multivariate circular and spherical data, it is relevant to test for independence among the components or subsets of components of the random vector. Again, a likelihood ratio test can be implemented by comparing the maximized log-likelihood of the considered joint model against the maximized log-likelihood of the corresponding model assuming a certain structure of stochastic independence among the components of the random vector. For example, in the case of testing for independence among the components of a random vector, i.e., under the null hypothesis that the joint density is the product of the univariate marginal densities, the likelihood ratio test statistic, Λ_{Ind}, is defined as follows:

$$\Lambda_{Ind} = -2 \left(\sum_{k=1}^{R} \hat{l}_{M_k} - \hat{l}_{\underline{M}} \right), \tag{10}$$

where $\hat{l}_{\underline{M}}$ is the maximized log-likelihood of the joint model with $\underline{M} = (M_1, \ldots, M_R)$ and $\hat{l}_{M_k}$ is the maximized log-likelihood for the k-th component of the random vector. The test statistic, Λ_{Ind}, asymptotically follows a chi-squared distribution with degrees of freedom $d.f. = 2 \prod_{k=1}^{R}(M_k + 1) - 2 - 2 \sum_{k=1}^{R} M_k$. In the case of spherical data, one can only test for independence among the latitude and longitude components of the spherical random vector.

When comparing K different populations, a homogeneity test is typically relevant. Again, a homogeneity likelihood ratio test can be constructed by comparing the maximized log-likelihood of the homogeneous model when fitting an NNTS (MNNTS or SNNTS) model to the pooled data from the K different populations to the sum of the maximized log-likelihoods of the NNTS (MNNTS or SNNTS) models fitted to the data of each of the K different populations being compared. Interested readers can consult Fernández-Durán and Gregorio-Domínguez (2010a) for the construction of a likelihood ratio homogeneity test in the case of circular data. The test statistic, Λ_{Hom}, asymptotically follows a chi-squared distribution. If the homogeneous model uses an $\underline{M} = (M_1, M_2, \ldots, M_R)$ then, the $d.f.$ of Λ_{Hom} is equal to $(K-1)\left(2\prod_{k=1}^{R}(M_k+1)-2\right)$.

3. The functions in CircNNTSR

The R (R Core Team 2016) package **CircNNTSR** (Fernández-Durán and Gregorio-Domínguez 2016) is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=CircNNTSR` and includes functions to perform maximum likelihood estimations, calculate and plot the density and distribution functions, and perform simulations of NNTS, MNNTS, and SNNTS models.

The names of functions for circular data begin with `nnts`. For example, the function `nntsdensity` calculates the density of an NNTS model with given c and M parameters. Similarly, the names of functions for spherical data begin with `snnts`, and those for multivariate circular data begin with `mnnts`. The package **CircNNTSR** also includes functions to consider grouped or incidence circular data in which only the total number of observations in different intervals of the random variable is provided. The names of functions for incidence data in the interval $(0, 2\pi]$ end with `interval0to2pi`, and those in the interval $(0, 1]$, which is commonly used when considering data on the occurrence of events in time, end with `interval0to1`.

The function that implements the Newton-like algorithm on a hypersphere to obtain the maximum likelihood estimates of the c parameters is named `nntsmanifoldnewtonestimation`, with an `s` and an `m` adding to the beginning of the name for the spherical and multivariate circular cases, respectively. For example, the following commands fit an NNTS circular model with $M = 3$ to data on the directions taken by 76 turtles after a treatment. The data were analyzed by Stephens (1969) and Fisher (1993, dataset B.3).

```
R> library("CircNNTSR")
R> data("Turtles_radians", package = "CircNNTSR")
R> cnnts <- nntsmanifoldnewtonestimation(data = Turtles_radians, M = 3,
+    iter = 1000)
R> cnnts

$cestimates
  k              cestimates
1 0   0.28645175+0.00000000i
2 1   0.11655438-0.12669303i
3 2  -0.14659000-0.16080633i
4 3   0.01079598+0.00065866i
```

```
$loglik
[1] -107.9374

$AIC
[1] 227.8749

$BIC
[1] 241.8593

$gradnormerror
[1] 2.466083e-08
```

The value of `M` represents the maximum number of modes in the data; the exploratory analysis of the data could give an indication of this value. Also, the degree of concentration of data around the modes should be considered. For highly concentrated data, large values of `M` are needed in order to get a good fit. The value of `iter` depends upon the complexity of the model, it should be large enough for the norm of the gradient to be near zero, i.e., within the desired error margin. The function `nntsmanifoldnewtonestimation` returns a list with a matrix containing the c parameter estimates (`cestimates`). The first column of the matrix contains the index of the c parameters. The other elements in the list are the values of the maximized log-likelihood (`loglik`), the Akaike's information criterion (`AIC`), the Bayesian information criterion (`BIC`), and the norm of the gradient (`gradnormerror`). The function `nntsmanifoldnewtonestimation` has two additional arguments, `initialpoint` and `cinitial`, that allow the user to begin iterations of the Newton algorithm from a different point. Generally, this option is used to check the convergence of the Newton algorithm from different initial points. The function `nntsrandominitial` can be used to randomly select a value for `cinitial`. For example, one could type

```
R> cnnts <- nntsmanifoldnewtonestimation(data = Turtles_radians, M = 3,
+     iter = 2000, initialpoint = TRUE, cinitial = nntsrandominitial(M = 3))
```

The following commands are used to plot the fitted NNTS model and the histogram of the data that are presented in Figure 1. Note that the values of the fitted NNTS density are equal at zero and 2π satisfying the periodicity condition for the density of a circular random variable.

```
R> hist(Turtles_radians, freq = FALSE, xlim = c(0, 2 * pi), ylim = c(0, 0.5),
+     breaks = seq(0, 2 * pi, (2 * pi) / 7))
R> nntsplot(cpars = cnnts$cestimates[, 2], M = 3, add = TRUE)
```

Similarly, the following commands fit and plot the fitted density, including the fitted univariate marginal densities, for a bivariate MNNTS model with $\underline{M} = (M_1, M_2) = (2, 2)$ for a dataset on the dihedral angles in a protein between three consecutive Alanine (Ala) amino acids. This dataset was constructed from the recommended July 2003 list of proteins via the algorithm in Hobohm, Scharf, Schneider, and Sander (1992) and Hobohm and Sander (1994). This algorithm selects a representative sample of proteins from the vast Protein Data Bank, PDB (Berman *et al.* 2000). The dataset contains 233 pairs of dihedral angles.
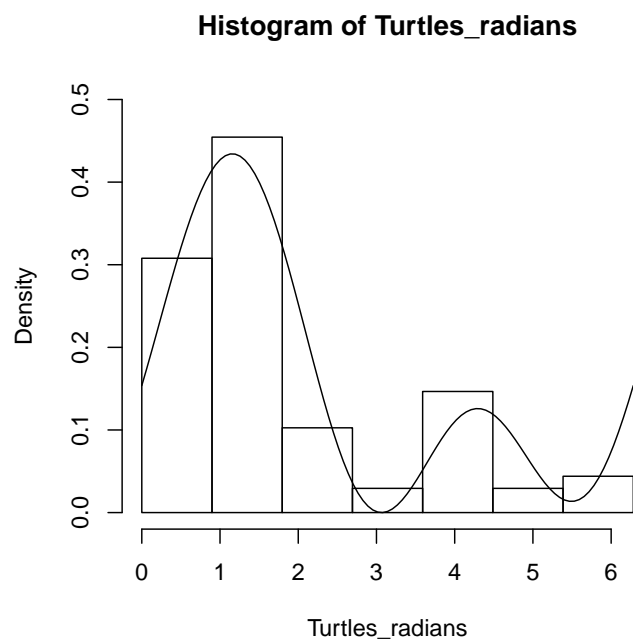
Histogram of Turtles_radians

Figure 1: Turtles data: Histogram and fitted NNTS model with $M = 3$ for data on the directions taken by 76 turtles after treatment.

```
R> data("ProteinsAAA", package = "CircNNTSR")
R> cmnnts <- mnntsmanifoldnewtonestimation(data = ProteinsAAA, M = c(2, 2),
+    R = 2, iter = 1000, initialpoint = TRUE, cinitial = mnntsrandominitial(
+    M = c(2, 2), R = 2))
R> cmnnts

$cestimates
  1 2               cestimates
1 0 0 0.04964464+0.00000000i
2 1 0 0.03477484+0.02866202i
3 2 0 0.01085742+0.03552719i
4 0 1 0.06465714-0.01033865i
5 1 1 0.05748308+0.04123364i
6 2 1 0.01093131+0.06455958i
7 0 2 0.03713285-0.00110292i
8 1 2 0.03829611+0.02375307i
9 2 2 0.01591254+0.04702533i

$loglik
[1] -451.9741

$AIC
[1] 935.9482

$BIC
```
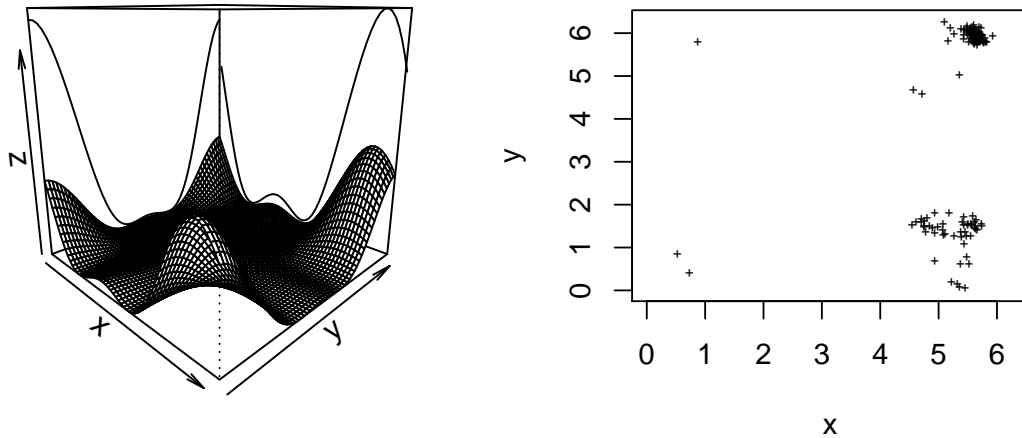
Figure 2: Dihedral angle data: Fitted bivariate MNNTS density indicating the univariate NNTS marginal densities as well and Ramachandran plot.

```
[1] 991.1648

$gradnormerror
[1] 1.354418e-16
```

The output of `mnntsmanifoldnewtonestimation` is very similar to the corresponding function in the univariate circular case. The first two columns of the element `cestimates` in the list contain the index of the estimated c parameters, and the third column contains the values of the estimates. The other elements of the list are the maximized log-likelihood value (`loglik`), the `AIC`, the `BIC`, and the norm of the gradient (`gradnormerror`).

The following command produces the plot of the fitted bivariate MNNTS density and the corresponding univariate marginal densities in the left plot in Figure 2. The x axis corresponds to the ϕ dihedral angle, and the y axis corresponds to the ψ dihedral angle. The Ramachandran plot, the two dimensional plot in Cartesian coordinates with angle ϕ along the horizontal axis and angle ψ along the vertical axis (Ramachandran, Ramakrishnan, and Sasisekharan 1963), is also included on the right side of Figure 2.

```
R> mnntsplotwithmarginals(cestimates = cmnnts$cestimates, M = c(2, 2),
+    theta = 45)
R> plot(ProteinsAAA, xlab = "x", ylab = "y", xlim = c(0, 2 * pi),
+    ylim = c(0, 2 * pi), cex = 0.5)
```

Finally, the following commands fit an SNNTS model with $\underline{M} = (M_1, M_2) = (3, 1)$ for a dataset on the measurements of magnetic remanence from specimens of red beds from the Bowen Basin, Queensland (Fisher, Lewis, and Embleton 1987, dataset B.5). The dataset contains 52 pairs of angles in declination-inclination coordinates that are transformed into longitude-latitude coordinates.

```
R> data("DataB5FisherSpherical", package = "CircNNTSR")
R> DataB5FisherSpherical[, 1] <- 360 - DataB5FisherSpherical[, 1]
```

```
R> DataB5FisherSpherical[, 2] <- 90 + DataB5FisherSpherical[, 2]
R> DataB5FisherSpherical2 <- DataB5FisherSpherical * (pi / 180)
R> M <- c(3, 1)
R> cpar <- rnorm(prod(M + 1)) + rnorm(prod(M + 1)) * complex(real = 0,
+    imaginary = 1)
R> cpar[1] <- Re(cpar[1])
R> cpar <- cpar / sqrt(sum(Mod(cpar) ^ 2))
R> csnnts<-snntsmanifoldnewtonestimation(data = DataB5FisherSpherical2,
+    M = c(3, 1), iter = 5000, initialpoint = TRUE, cinitial = cpar)
R> csnnts


$cestimates
  k1 k2          cestimates
1  0  0   0.4961416+0.0000000i
5  0  1  -0.0476352-0.4945939i
2  1  0   0.1615553-0.0758675i
6  1  1  -0.1403093-0.1210623i
3  2  0   0.1878969-0.2490072i
7  2  1  -0.2674327-0.1944443i
4  3  0   0.1601738+0.3147032i
8  3  1   0.2811328-0.1743109i

$loglik
[1] -112.2896

$AIC
[1] 252.5792

$BIC
[1] 279.8966

$gradnormerror
[1] 2.158877e-16
```

When using an algorithm to implement the Newton-like algorithm on the hypersphere, it is important to test different initial points to avoid local maxima. These initial points can be selected randomly using the function `nntsrandominitial`. If several random initial points result in the same maximum, this increases the certainty that the algorithm did not get trapped in local maxima. In our experience, for many datasets that we have analyzed, one gets consistently the same maximum when using different random initial points. Difficult datasets in which the algorithm found different maxima were detected when running the algorithm from five to ten different random initial points. For these cases, the algorithm must be executed more times from different random initial points until one gets many replicates resulting in the same maximum. In the next section, we present examples with actual datasets using the package **CircNNTSR**.

4. Examples

4.1. Circular univariate data: NNTS

We consider data on the time of occurrence of earthquakes of intensity greater than 6.0 on the Richter scale with the epicenter occurring on the coast of the Pacific Ocean in Mexico from 1920 to 2002 (Fernández-Durán and Gregorio-Domínguez 2016, `EarthquakesPacificMexico-gt6` dataset). There are a total of 241 observations. We apply a uniformity test by fitting NNTS models with $M = 0, 1, \ldots, 6$. Table 1 includes the values of the maximized log-likelihood, $\hat{l}$, the likelihood ratio uniformity test statistic, Λ_{Unif}, the degrees of freedom, $d.f.$, and the p value for the uniformity test. From the results in Table 1, the null hypothesis of uniformity is not rejected at the significance level of 5% although for the tests with $M = 2, 4, 6$ the p values are close to 5%. By applying other tests of uniformity included in the **circular** package (Agostinelli and Lund 2013), the null hypothesis of uniformity is rejected by the tests of Kuiper and Watson but not rejected by the Rayleigh's test when applied to the `EarthquakesPacificMexicogt6` dataset at a 5% significance level. Rejection of the null hypothesis of uniformity could be related to the existence of aftershocks of intensity greater than 6.0 on the Richter scale.

By considering only the 76 earthquakes of intensity greater than 7 on the Richter scale (Fernández-Durán and Gregorio-Domínguez 2016, dataset `EarthquakesPacificMexicogt7`), Table 2 includes the results for tests of uniformity; the p values in Table 2 are greater than those in Table 1, thus clearly not rejecting the null hypothesis of uniformity. For the case of earthquakes of intensity greater than 7.0 on the Richter scale where the existence of large aftershocks is very low, the uniformity tests of Kuiper, Rayleigh and Watson also do not reject the null hypothesis of uniformity. These results support the idea of using a homogeneous Poisson process to model the times of occurrence of high-intensity earthquakes.

We apply a homogeneity test to compare the time of occurrence (starting times) of hurricanes in the Gulf of Mexico for the 1951–1970 and 1971–2008 periods. There are a total of 196 and 417 observations for the 1951–1970 period (Fernández-Durán and Gregorio-Domínguez 2016, dataset `HurricanesGulfofMexico1951to1970`) and 1971–2008 period (Fernández-Durán and Gregorio-Domínguez 2016, dataset `HurricanesGulfofMexico1971to2008`), respectively. Table 3 includes the maximized log-likelihoods for the data in the two periods and, for $M = 0, 1, \ldots, 6$, the value of the homogeneity test statistic, Λ_{Hom}, the degrees of freedom, and the corresponding p values. From the results in Table 3, the null hypothesis of homogeneity is

M	$\hat{l}$	Λ_{Unif}	$d.f.$	χ^2 p value
0	-442.93			
1	-442.33	1.1954	2	0.5501
2	-438.37	9.1168	4	0.0582
3	-438.18	9.4891	6	0.1479
4	-435.50	14.8545	8	0.0620
5	-435.26	15.3415	10	0.1201
6	-432.75	20.3489	12	0.0608

Table 1: Uniformity test results for earthquakes occurring in the 1920–2002 period with intensity greater than 6 on the Richter scale.

M	$\hat{l}$	Λ_{Unif}	d.f.	χ^2 p value
0	-139.68			
1	-139.13	1.0949	2	0.5784
2	-138.89	1.5830	4	0.8118
3	-138.47	2.4164	6	0.8777
4	-135.48	8.3901	8	0.3963
5	-135.41	8.5328	10	0.5769
6	-133.67	12.0214	12	0.4440

Table 2: Uniformity test results for earthquakes occurring in the 1920–2002 period with intensity greater than 7 on the Richter scale.

M_0	$\hat{l}$ 1951–1970	$\hat{l}$ 1971–2008	$\hat{l}$ 1951–2008	Λ_{Hom}	d.f.	χ^2 p value
0	-360.22	-766.39	-1126.62		0	
1	-250.95	-530.77	-781.73	0.0133	2	0.9934
2	-213.70	-456.59	-671.30	2.0180	4	0.7325
3	-203.82	-431.76	-636.81	2.4552	6	0.8734
4	-196.95	-424.79	-626.34	9.2073	8	0.3251
5	-191.22	-419.87	-615.73	9.2985	10	0.5040
6	-191.11	-415.64	-612.27	11.0452	12	0.5250
7	-190.59	-415.34	-611.77	11.6920	14	0.6310
8	-189.71	-414.97	-611.65	13.9427	16	0.6030
9	-189.35	-413.38	-610.63	15.8097	18	0.6058
10	-188.19	-413.37	-610.37	17.6220	20	0.6123

Table 3: Homogeneity test results for the times of occurrence of hurricanes in two periods: 1951–1970 and 1971–2008.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Suicides	318	335	370	360	428	367	366	386	390	342	331	320
%	7.37	7.77	8.58	8.35	9.92	8.51	8.49	8.95	9.04	7.93	7.67	7.42
Homicides	776	776	819	859	822	889	817	792	837	814	825	881
%	7.83	7.83	8.27	8.67	8.30	8.97	8.25	7.99	8.45	8.22	8.33	8.89

Table 4: Monthly number of suicides and homicides in Mexico during 2005.

not rejected, indicating that there is no evidence to suggest that the occurrence patterns for 1951–1970 are different from those for 1971–2008.

As an example of a homogeneity test for incidence data, we test the monthly number of suicides (Fernández-Durán and Gregorio-Domínguez 2016, dataset `SuicidesMexico2005`) and homicides (Fernández-Durán and Gregorio-Domínguez 2016, dataset `HomicidesMexico2005`) in Mexico during 2005 for homogeneity. Table 4 includes the data reported by the Mexican National Statistical Agency INEGI.

Table 5 includes the results of the homogeneity test for suicides and homicides in which the null hypothesis of homogeneity is rejected at the 5% significance level. To fit an NNTS model to the monthly number of homicides or suicides, the next commands using functions of **CircNNTSR** are used to fit an NNTS model with $M = 1$ to the incidence (grouped) data

M	$\hat{l}$ suicides 2005	$\hat{l}$ homicides 2005	$\hat{l}$ suicides and homicides 2005	Λ_{Hom}	$d.f.$	χ^2 p value
0	-10717.05	-24620.57	-35337.62		0	
1	-10707.97	-24620.20	-35333.37	10.41	2	0.0055
2	-10706.07	-24618.29	-35331.64	14.58	4	0.0057
3	-10705.13	-24618.35	-35331.44	15.92	6	0.0142
4	-10702.36	-24614.96	-35327.72	20.78	8	0.0078
5	-10701.89	-24611.23	-35326.77	27.31	10	0.0023
6	-10701.56	-24609.14	-35323.52	25.64	12	0.0121

Table 5: Homogeneity test results for the monthly number of suicides and homicides in Mexico during 2005.

in the number of homicides in Mexico during 2005.

```
R> data("HomicidesMexico2005", package = "CircNNTSR")
R> data("SuicidesMexico2005", package = "CircNNTSR")
R> months <- c(0, 31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
R> monthsac <- cumsum(months) / 365
R> monthsac[13] <- 1 - 0.00000001
R> reshom1 <- nntsmanifoldnewtonestimationinterval0to1(
+    data = HomicidesMexico2005, cutpoints = monthsac, subintervals = 12,
+    M = 1, iter = 10000)
R> reshom1

$cestimates
  k              cestimates
1 0   0.398934754+0.000000000i
2 1  -0.001401083-0.002010551i

$loglik
        [,1]
[1,] -24620.2

$gradnormerror
[1] 0
```

Figure 3 includes the histograms for the monthly number of suicides and homicides in Mexico during 2005. In Figure 3, the fitted density function of the best NNTS model in terms of the BIC is superimposed in each histogram. The results of the homogeneity test in Table 5 confirm the results of previous studies that the occurrence of suicides peaks during the year whereas the occurrence of homicides is distributed uniformly throughout the year.

4.2. Multivariate circular data: MNNTS

The pollution monitoring network of Mexico's Central Valley consists of 36 monitoring stations located in Mexico City and neighboring states that measure pollution variables, such as
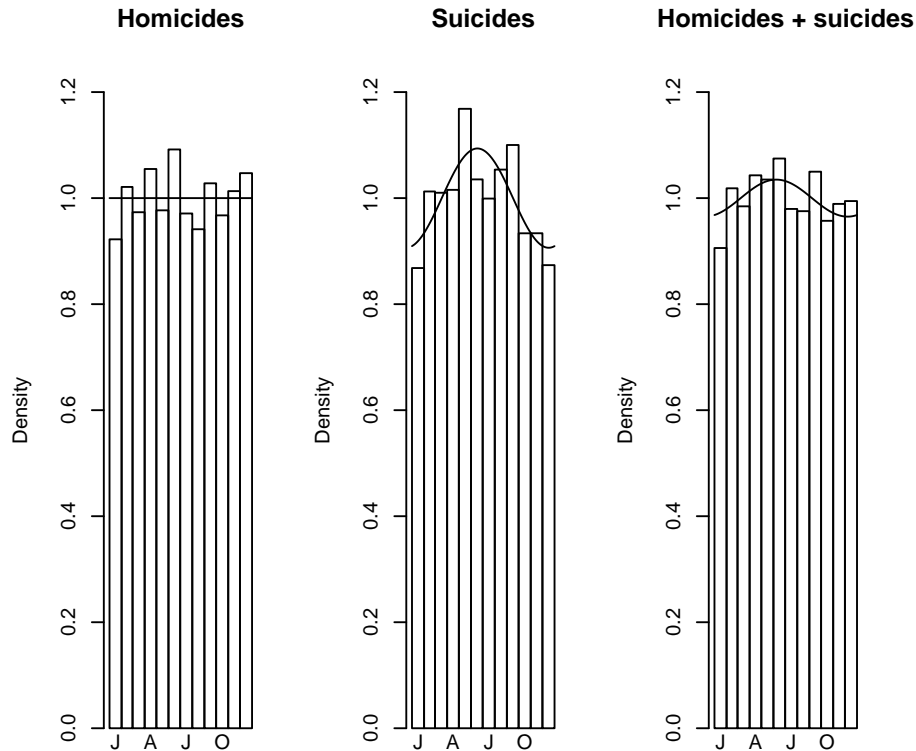
Figure 3: Monthly homicides and suicides in Mexico during 2005: Histograms and superimposed estimated densities from the best BIC NNTS models.

ozone (O_3), carbon monoxide (CO), and nitrogen oxides (NO_x), along with such meteorological variables as wind direction and speed, temperature, and humidity. In this application, we consider wind directions registered at the monitoring stations of *San Agustín* located in the north, *Pedregal* in the southwest and *Hangares* in the southeast at 14:00 h on the days between January 1, 1993 and February 29, 2000. There are a total of 1,682 observations in the **CircNNTSR** dataset `WindDirectionsTrivariate`.

We fitted trivariate MNNTS models with M parameters up to $M_1 = 3$, $M_2 = 3$, and $M_3 = 3$ (see Table 6). The best model according to BIC is the one with $M_1 = 3$, $M_2 = 2$, and $M_3 = 3$.

A model that has the same orders $M_1 = 3$, $M_2 = 2$, and $M_3 = 3$, but assumes independence among the three components has a much larger BIC than the dependent joint MNNTS model. Table 7 provide the maximized log-likelihood values of the univariate marginal models. All of the likelihood ratio tests for different combinations of M_1, M_2 and M_3 clearly reject the null hypothesis of independence, with p values less than 10^{-6}.

As another example of multivariate circular data, we simulated 200 realizations from a bivariate circular uniform distribution with independent components and fitted MNNTS to these data with $M_1 = 0, 1, \ldots, 3$ and $M_2 = 0, 1, \ldots, 3$. This dataset is included in **CircNNTSR** with the name `DataUniformBivariate200obs`. As expected, the best model in terms of BIC is the one with $M_1 = 0$ and $M_2 = 0$, which corresponds to a bivariate circular uniform distribution. Table 8 presents the bivariate results, and Table 9 presents the univariate results. For all of the cases considered, we do not reject the null hypothesis of independence, with p values

M_1	M_2	M_3	$\hat{l}$	$f.p.$	BIC	M_1	M_2	M_3	$\hat{l}$	$f.p.$	BIC
0	0	0	-9273.93	0	18547.86	2	0	0	-8901.12	4	17831.95
0	0	1	-9048.22	2	18111.29	2	0	1	-8408.70	10	16891.69
0	0	2	-8896.18	4	17822.07	2	0	2	-8204.25	16	16527.33
0	0	3	-8861.65	6	17767.86	2	0	3	-8150.22	22	16463.84
0	1	0	-9192.30	2	18399.46	2	1	0	-8793.04	10	17660.35
0	1	1	-8938.57	6	17921.70	2	1	1	-8278.96	22	16721.32
0	1	2	-8781.49	10	17637.26	2	1	2	-8076.01	34	16404.56
0	1	3	-8763.51	14	17631.00	2	1	3	-7983.63	46	16308.93
0	2	0	-9067.42	4	18164.54	2	2	0	-8634.30	16	17387.45
0	2	1	-8754.63	10	17583.54	2	2	1	-8094.89	34	16442.33
0	2	2	-8574.57	16	17267.98	2	2	2	-7855.90	52	16098.04
0	2	3	-8537.61	22	17238.63	2	2	3	-7757.95	70	16035.84
0	3	0	-9033.57	6	18111.71	2	3	0	-8574.87	22	17313.14
0	3	1	-8743.33	14	17590.64	2	3	1	-8037.22	46	16416.12
0	3	2	-8530.44	22	17224.30	2	3	2	-7813.67	70	16147.28
0	3	3	-8533.71	30	17290.26	2	3	3	-7683.75	94	16065.70
1	0	0	-9118.72	2	18252.30	3	0	0	-8841.18	6	17726.93
1	0	1	-8590.65	6	17225.86	3	0	1	-8310.83	14	16725.64
1	0	2	-8422.29	10	16918.86	3	0	2	-8104.54	22	16372.50
1	0	3	-8364.70	14	16833.39	3	0	3	-8040.25	30	16303.34
1	1	0	-9014.75	6	18074.06	3	1	0	-8732.45	14	17568.89
1	1	1	-8488.48	14	17080.95	3	1	1	-8171.15	30	16565.14
1	1	2	-8361.19	22	16885.80	3	1	2	-7993.24	46	16328.17
1	1	3	-8253.79	30	16730.41	3	1	3	-7890.86	62	16242.25
1	2	0	-8849.93	10	17774.13	3	2	0	-8572.16	22	17307.73
1	2	1	-8276.58	22	16716.57	3	2	1	-7957.50	46	16256.67
1	2	2	-8105.41	34	16463.36	3	2	2	-7740.43	70	16000.80
1	2	3	-7986.71	46	16315.10	3	2	3	-7634.01	94	15966.23
1	3	0	-8822.06	14	17748.10	3	3	0	-8521.00	30	17264.83
1	3	1	-8264.95	30	16752.74	3	3	1	-7926.23	62	16312.98
1	3	2	-8034.50	46	16410.68	3	3	2	-7687.76	94	16073.72
1	3	3	-7940.16	62	16340.85	3	3	3	-7562.57	126	16061.03

Table 6: Wind direction data: Value of the maximized log-likelihood, number of free parameters, and BIC for each MNNTS model considered. $f.p.$ denotes the number of free parameters.

M	$f.p.$	$\hat{l}_1$	BIC_1	$\hat{l}_2$	BIC_2	$\hat{l}_3$	BIC_3
0	0	-3091.31	6182.62	-3091.31	6182.62	-3091.31	6182.62
1	2	-2936.11	5887.07	-3009.69	6034.23	-2865.60	5746.05
2	4	-2718.50	5466.72	-2884.80	5799.30	-2713.56	5456.83
3	6	-2658.56	5361.69	-2850.95	5746.47	-2679.03	5402.62

Table 7: Wind direction data: Maximized log-likelihood and BIC values of the univariate marginal NNTS models used to construct the test for independence.

M_1	M_2	$f.p.$	$\hat{l}$	BIC	Λ_{Ind}	$d.f.$	χ^2 p value
0	0	0	-735.15	1470.30		0	
0	1	2	-735.03	1480.66		0	
0	2	4	-734.54	1490.27		0	
0	3	6	-733.54	1498.86		0	
1	0	2	-735.02	1480.64		0	
1	1	6	-733.94	1499.67	1.9154	2	0.3838
1	2	10	-732.63	1518.23	3.5600	4	0.4688
1	3	14	-730.80	1535.78	5.2021	6	0.5182
2	0	4	-734.75	1490.70		0	
2	1	10	-733.13	1519.25	2.9952	4	0.5586
2	2	16	-731.23	1547.24	5.8119	8	0.6683
2	3	22	-727.16	1570.87	11.9656	12	0.4484
3	0	6	-734.08	1499.96		0	
3	1	14	-731.28	1536.74	5.3628	6	0.4982
3	2	22	-728.60	1573.76	9.7472	12	0.6381
3	3	30	-723.17	1605.30	18.5936	18	0.4172

Table 8: Simulated uniform data: BIC, likelihood-ratio test statistic for independence (Λ_{Ind}), and p value for each MNNTS model considered.

M	$f.p.$	$\hat{l}_1$	BIC_1	$\hat{l}_2$	BIC_2
0	0	-367.58	735.15	-367.58	735.15
1	2	-367.44	745.49	-367.45	745.51
2	4	-367.18	755.55	-366.96	755.12
3	6	-366.51	764.81	-365.96	763.71

Table 9: Simulated uniform data: Maximized log-likelihood and BIC values of the marginal NNTS models used to construct the test for independence.

greater than 0.05.

The following commands can be used to generate 500 realizations from an MNNTS model with $\underline{M} = (2, 5, 1)$ and randomly selected c parameters (`cparsim`):

```
R> cparsim <- mnntsrandominitial(M = c(2, 5, 1), R = 3)
R> simulations <- mnntssimulation(nsim = 500, cpar = cparsim, M = c(2, 5, 1),
+    R = 3)
```

The simulation algorithm is based on the acceptance-rejection method of simulation as suggested by Fernández-Durán and Gregorio-Domínguez (2014b). Since the Marsenne-Twister pseudorandom number generator is the default in R, when a massive number of realizations from an (MS)NNTS model is required, the proposed acceptance-rejection algorithm can be implemented in parallel to create independent random streams in different machines by applying the dynamic creation of pseudorandom number generators proposed by Matsumoto and Nishimura (2000).
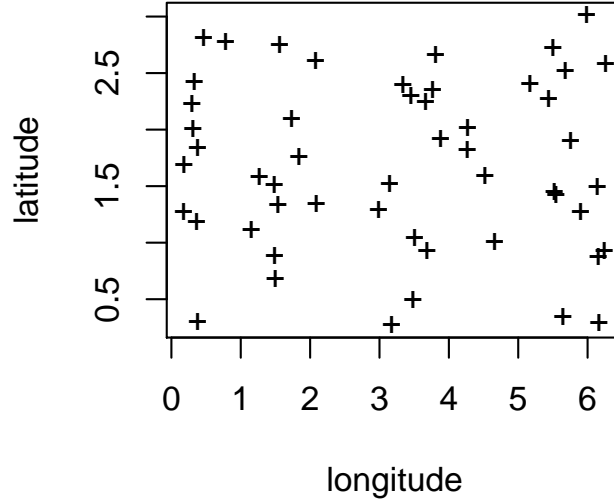
Figure 4: Scatterplot of measurements of magnetic remanence from 52 specimens of red beds from the Bowen Basin, Queensland (see Fisher *et al.* 1987, dataset B.5).

M_1	M_2	$\hat{l}$	Λ_{Unif}	*d.f.*	χ^2 *p* value
0	0	-151.11			
0	1	-122.58	57.0689	2	0.0000
0	2	-120.39	61.4381	4	0.0000
1	0	-149.59	3.0523	2	0.2174
1	1	-121.03	60.1698	6	0.0000
1	2	-118.72	64.7829	10	0.0000
2	0	-147.17	7.8809	4	0.0960
2	1	-118.61	65.0167	10	0.0000
2	2	-115.94	70.3389	16	0.0000

Table 10: Uniformity test results for magnetic remanence measurements in red beds.

4.3. Spherical data: SNNTS

Figure 4 includes the scatterplot for the 52 measurements of magnetic remanence from specimens of red beds from the Bowen Basin, Queensland (Fisher *et al.* 1987, dataset B.5.). Based on Figure 4, it is relevant to test for spherical uniformity using the **CircNNTSR** dataset `DataB5FisherSpherical`.

Table 10 includes the values of the maximized log-likelihoods for SNNTS models with the combinations of $M_1 = 0, 1, 2, 3$ and $M_2 = 0, 1, 2, 3$ and the values of the test statistic for uniformity, Λ_{Unif}, and asymptotic *p* value. From these results, the null hypothesis of uniformity is rejected at the 5% significance level for all cases except $\underline{M} = (1, 0)$ and $\underline{M} = (2, 0)$, which correspond to radially symmetric distributions; although these models gave a very poor fit relative to others, except $\underline{M} = (0, 0)$.

5. Summary

The R package **CircNNTSR** includes functions for the statistical analysis of circular, multivariate circular and spherical data using models based on nonnegative trigonometric sums. Since the parameter space of NNTS, MNNTS and SNNTS models corresponds to the surface of a hypersphere, maximum likelihood estimation is implemented by a Newton-like algorithm on manifolds. (MS)NNTS models are useful to analyze datasets that present skewness and multiple modes but that are not highly concentrated around the modes. For extremely highly concentrated datasets one can consider more parsimonious models such as those based on, for example, the von Mises distribution and its generalizations. Comparisons in terms of AIC and BIC criteria among (MS)NNTS models and alternative models, such as those included in the **circular** (Agostinelli and Lund 2013) and **movMF** (Hornik and Grün 2014) R packages, can be implemented to assess the most adequate model for a particular dataset. A future extension to the **CircNNTSR** package under current development is the inclusion of R functions for the statistical analysis of (MS)NNTS models from a Bayesian perspective considering Markov chain Monte Carlo (MCMC) algorithms.

Acknowledgments

References

Agostinelli C, Lund U (2013). *R Package* **circular**: *Circular Statistics (Version 0.4-7)*. CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA. URL https://CRAN.R-project.org/package=circular.

Berman H, Westbrook J, Feng Z, Gilliand G, Bhat T, Weissing H, Shyndialov I, Bourne P (2000). "The Protein Data Bank." *Nucleic Acids Research*, **28**(1), 235–242. doi:10.1093/nar/28.1.235.

Fejér L (1916). "Über trigonometrische Polynome." *Journal für die reine und angewandte Mathematik*, **1916**(146), 53–82. doi:10.1515/crll.1916.146.53.

Fernández-Durán JJ (2004). "Circular Distributions Based on Nonnegative Trigonometric Sums." *Biometrics*, **60**(2), 499–503. doi:10.1111/j.0006-341x.2004.00195.x.

Fernández-Durán JJ, Gregorio-Domínguez MM (2010a). "A Likelihood-Ratio Test for Homogeneity in Circular Data." *Journal of Biometrics & Biostatistics*, **1**(3), 107. doi:10.4172/2155-6180.1000107.

Fernández-Durán JJ, Gregorio-Domínguez MM (2010b). "Maximum Likelihood Estimation of Nonnegative Trigonometric Sums Models Using a Newton-Like Algorithm on Manifolds." *Electronic Journal of Statistics*, **4**, 1402–1410. doi:10.1214/10-ejs587.

Fernández-Durán JJ, Gregorio-Domínguez MM (2014a). "Distributions for Spherical Data Based on Nonnegative Trigonometric Sums." *Statistical Papers*, **55**(4), 983–1000. `doi:10.1007/s00362-013-0547-5`.

Fernández-Durán JJ, Gregorio-Domínguez MM (2014b). "Modeling Angles in Proteins and Circular Genomes Using Multivariate Angular Distributions Based on Nonnegative Trigonometric Sums." *Statistical Applications in Genetics and Molecular Biology*, **13**(1), 1–18. `doi:10.1515/sagmb-2012-0012`.

Fernández-Durán JJ, Gregorio-Domínguez MM (2014c). "Testing for Seasonality Using Circular Distributions Based on Nonnegative Trigonometric Sums as Alternative Hypotheses." *Statistical Methods in Medical Research*, **23**(3), 279–292. `doi:10.1177/0962280211411531`.

Fernández-Durán JJ, Gregorio-Domínguez MM (2016). **CircNNTSR**: *An* R *Package for the Statistical Analysis of Circular Data Using Nonnegative Trigonometric Sums (NNTS) Models*. R package version 2.2, URL `https://CRAN.R-project.org/package=CircNNTSR`.

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

Fisher N, Lewis T, Embleton B (1987). *Statistical Analysis of Spherical Data*. Cambridge University Press.

Hobohm U, Sander C (1994). "Enlarged Representative Set of Protein Structures." *Protein Science*, **3**(3), 522–524. `doi:10.1002/pro.5560030317`.

Hobohm U, Scharf M, Schneider R, Sander C (1992). "Selection of a Representative Set of Structures from the Brookhaven Protein Data Bank." *Protein Science*, **1**(3), 409–417. `doi:10.1002/pro.5560010313`.

Hornik K, Grün B (2014). "**movMF**: An R Package for Fitting Mixtures of Von Mises-Fisher Distributions." *Journal of Statistical Software*, **58**(10), 1–31. `doi:10.18637/jss.v058.i10`.

Linkie M, Ridout M (2011). "Assessing Tiger-Prey Interactions in Sumatran Rainforests." *Journal of Zoology*, **284**(3), 224–229. `doi:10.1111/j.1469-7998.2011.00801.x`.

Lynam A, Jenks K, Tantipisanuh N, Chutipong W, Ngoprasert D, Gales G, Steinmetz R, Sukmasuang R, Bhumpakphan N, Grassman Jr LI, Cutter P, Kitamura S, Reed D, Baker M, McShea W, Songsasen N, Leimgruber P (2013). "Terrestrial Activity Patterns of Wild Cats from Camera-Trapping." *The Raffles Bulletin of Zoology*, **61**, 407–415.

Matsumoto M, Nishimura T (2000). "Dynamic Creation of Pseudorandom Number Generators." In *Monte Carlo and Quasi-Monte Carlo Methods 1998*, pp. 56–69. Springer-Verlag.

Ramachandran G, Ramakrishnan C, Sasisekharan V (1963). "Stereochemistry of Polypeptide Chain Configurations." *Journal of Molecular Biology*, **7**(1), 95–99. `doi:10.1016/s0022-2836(63)80023-6`.

R Core Team (2016). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ridout M, Linkie M (2009). "Estimating Overlap of Daily Activity Patterns from Camera Trap Data." *Journal of Agricultural, Biological and Environmental Statistics*, **14**(3), 322–337. `doi:10.1198/jabes.2009.08038`.

Stephens M (1969). "Techniques for Directional Data." *Technical Report 150*, Department of Statistics, Stanford University, Stanford, CA.

Affiliation:

Juan José Fernández Durán
School of Business
Instituto Tecnológico Autónomo de México
Río Hondo No. 1, Col. Progreso Tizapán, C.P. 01080, México D.F., México
E-mail: jfdez@itam.mx

María Mercedes Gregorio Domínguez
Department of Actuarial Science
Instituto Tecnológico Autónomo de México
Río Hondo No. 1, Col. Progreso Tizapán, C.P. 01080, México D.F., México
E-mail: mercedes@itam.mx