Circular Regression Models and Regression Trees with circtree

Moritz N. Lang
Universität Innsbruck

Abstract

The **circtree** package provides functions for maximum likelihood estimation of circular regression models employing a von Mises distribution. Additionally, a distribution tree can be fitted for a circular response employing the von Mises distribution and using the covariates as potential splitting variables. For both approaches suitable standard methods are provided to print the fitted models, and compute predictions and inference.

Keywords: regression, distribution tree, circular response, von Mises distribution, R.

1. Introduction

Circular response variables occur in a variety of applications and subject areas. E.g., gun crime data on a 24-hour scale is analysed in the social-economics, animal orientation or gene-structure analysis are often subject of examination in biology, and wind data is one of the most important weather variables in meteorology.

To represent different circular responses the von Mises distribution is employed. It is also known as the circular normal distribution, and is a special case of the von Mises-Fisher distribution on the N-dimensional sphere:

$$f(x \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \, e^{\kappa \cos(x-\mu)}, \tag{1}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order 0.

The **circtree** package provides functions to fit circular regression models by maximum likelihood estimation and to fit distribution trees for a circular response employing potential covariates as splitting variables. For both methods, a convenient formula interface and standard methods for analysis and prediction are available. For the illustration of the von Mises distribution, an interactive shiny app is additionally provided. To employ the von Mises distribution as circular response in other packages, families for **bamlss** and **disttree** are exported with necessary functions for the computation of scores and the Hessian matrix.

The outline of the paper is as follows. Section 3 describes the fitting of the circular regression model, and Section 3 presents methods for distribution trees for a circular response. For both methods different R implementations are illustrated by using artificial data. In the end of the paper, Section 4 provides a very brief summary of the two approaches.

2. Circular regression models

For circular regression models the response is assumed to follow a von Mises distribution VM as defined in Equation 1:

$$y \sim VM \tag{2}$$

The location parameter μ and the concentration parameter κ are assumed to be linked to the covariates $\mathbf{x} = (x_1, x_2, \ldots)^\top$ and $\mathbf{z} = (z_1, z_2, \ldots)^\top$:

$$\mu = \alpha_0 + g^{-1}(\mathbf{x}^\top \beta) \tag{3}$$

$$\kappa = h^{-1}(\gamma_0 + \mathbf{z}^\top \gamma), \tag{4}$$

where $\beta = (\beta_1, \beta_2, \ldots)^\top$ and $\gamma = (\gamma_1, \gamma_2, \ldots)^\top$ are the slope coefficients and α_0 and γ_0 the intercepts, respectively (Fisher and Lee 1992). The link functions $g(\cdot) : \mathbb{R} \mapsto (-\pi, \pi)$ and $h(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ are monotonic and twice differentiable functions. For the concentration parameter κ the logarithm function is typically employed (i.e., $h^\top(\cdot) = \exp(\cdot)$). For the location parameter μ the 'tan-half' link is a well suited function restricting the values to $(-\pi, \pi)$:

$$g^\top(\cdot) = 2 \arctan(\cdot). \tag{5}$$

The offset parameter α_0 outside of the inverse link function of the predictors performs a simple rotation of the response. To restrict the parameter α_0 to $(-\pi, \pi)$ we also apply the inverse link function $g^\top(\cdot)$ to it. Therefore, μ can theoretically take values between -2π and $+2\pi$, but has still a restricted range of 2π.

A circular regression model by maximum likelihood estimation can be fitted with the `circmax()` function provided by the **circtree** package. This function provides a standard formula interface with arguments like formula, data, subset, etc. It first sets up the likelihood function, gradients and Hessian matrix and uses `optim()` to maximize the von Mises likelihood. For the S3 return object various standard methods are available.

```
circmax(formula, data, subset, na.action, model = TRUE, y = TRUE,
        x = FALSE, control = circmax_control(...), ...)
```

Here `formula`, `data`, `subset`, and `na.action` have their standard model frame meanings (e.g., Chambers and Hastie 1992). However, as provided in the **Formula** package (Zeileis and Croissant 2010) `formula` can have two parts separated by '`|`' where the first part defines the location model and the second part the concentration model. E.g., with `y ~ x1 + x2 | z1 + z2` the location model is specified by `y ~ x1 + x2` and the concentration model by `~ z1 + z2`.

The maximum likelihood estimation is carried out with the R function `optim()` using control options specified in `circmax_control()`. By default the `"Nelder-Mead"` method is applied neglecting provided gradients. If no starting values are supplied, a closed form maximum likelihood estimator is applied for the starting values for the intercept of the location part. For the intercept of the concentration part, by default a Newton Fourier method is employed. The starting values for the regression coefficients in the location and concentration model are set by default to zero. The parameters `model`, `y`, and `x` specify whether the model frame, response, or model matrix should be returned.

The following example illustrates the function calls for the circular regression model for a artificial data set. First the **circtree** package is loaded and 1000 simulated observations are created by the function `circmax_simulate()` with location coefficients `beta` 3, 5, and 2 and concentration coefficients `gamma` 3 and 3.

```
R> library("circtree")
R> sdat <- circmax_simulate(n = 1000, beta = c(3, 5, 2), gamma = c(3, 3))
R> head(sdat)

           x1          x2         x3        y
1  0.04704414 -0.13235907  0.2268724 2.504737
2 -0.06614717  0.12410245  0.1715144 2.256686
3 -0.06232476 -0.05038766 -0.1147274 1.774435
4 -0.46046913 -0.55198373  0.2755990 6.137801
5 -0.03417521  0.10543545 -0.1150955 3.212166
6  0.02805565  0.04382372 -0.1185750 2.825178
```

We fit a circular regression by maximum likelihood employing the covariates x1, x2 for the location model and the covariate x3 for the concentration model. The results show that the fitted coefficients are quite near to the real values. The fitted model has a log-likelihood of 45.94 with 5 degree of freedom.

```
R> m.circmax <- circmax(y ~ x1 + x2 | x3, data = sdat)
R> print(m.circmax)

Maximum likelihood estimation for the von Mises distribution

Coefficients (location model with tanhalf link):
(Intercept)           x1           x2
      2.989        5.059        2.037

Coefficients (concentration model (density kappa) with log link):
(Intercept)           x3
      2.951        3.143

Log-likelihood: 45.94
Df: 5
```

3. Distribution trees for a von Mises distribution

As an alternative approach, a distribution tree for a circular response employing a von Mises distribution can be fitted with the `circtree()` function. This is a wrapper function for the `mob()` function provided in the **partykit** package (Zeileis, Hothorn, and Hornik 2008; Hothorn and Zeileis 2015). A fitting function `circfit()` for the parameter estimation on the given data is given in the **circtree** package so that MOB algorithm can employ all information needed for parameter instability tests and partitioning.

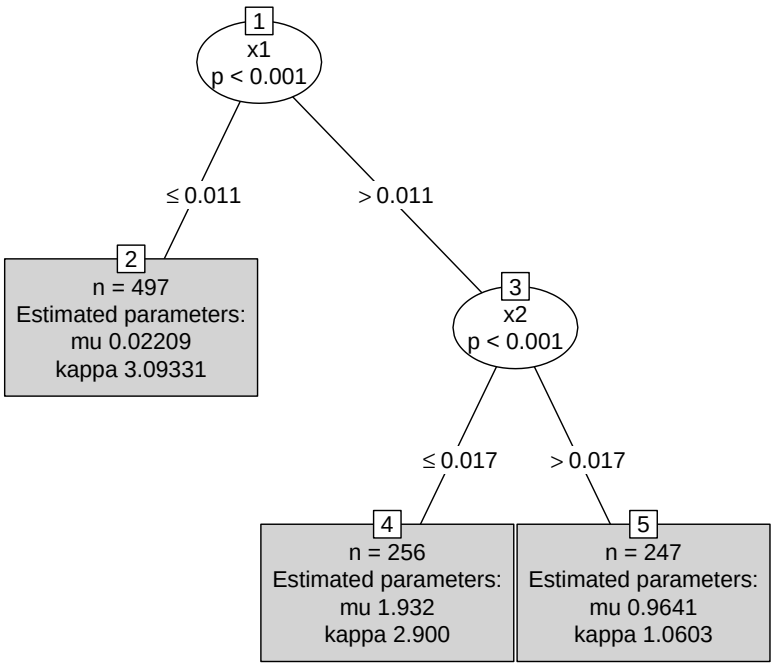

Figure 1: Fitted distribution tree for a circular response employing the von Mises distribution.

Both `circtree()` and `circfit()` support standard interfaces for e.g., formula, data, and subset arguments. `circtree()` first performs some intern checks, set ups the formula and the control arguments and than calls the `mob()` function within the **partykit** employing the fitting function `circfit()` for a circular response. The `circfit()` function provides the log-likelihood, score and hessian function for the von Mises distribution and performs the fitting of the distribution parameters. The `circfit()` function can be also called for distribution parameter fitting by itself. Therefore, various standard methods are provided for the S3 return objects of the `circtree()` and `circfit()` function calls.

```
circtree(formula, data, start, subset, na.action, weights, offset,
         control = partykit::mob_control(),
         fit_control = circfit_control(...), ...)
```

Here `formula`, `data`, `subset`, `na.action`, `weights`, and `offset` have their standard model frame meanings as described in Section 3. A list of control options for the `mob()` function can be set up by the `mob_control()` function, including options for pruning.

The distribution parameter estimation is carried out by maximum likelihood estimation. The location parameter is calculated by a closed form maximum likelihood estimator. For the concentration parameter a Newton Fourier method is by default employed controlled via the `fit_control()` function. Alternatively, a uniroot provides a safe estimation option and a method introduced by Banerjee, Dhillon, Ghosh, and Sra (2005) provides a quick

approximation of the concentration parameter. The starting values `start` are currently not used for the parameter estimation.

As in Section 3, the function calls for the regression tree employing a von Mises distribution are illustrated employing an artificial data set. First the **circtree** package is loaded and 1000 simulated observations are created by the function `circtree_simulate()`. We generate three groups with location parameters `mu` 0, 2, and 1 and concentration parameters `kappa` 3, 3, and 1, respectively.

```
R> library("circtree")
R> sdat <- circtree_simulate(n = 1000, mu = c(0, 2, 1), kappa = c(3, 3, 1))
R> head(sdat)
```

```
           x1          x2 group mu kappa         y
1  0.18596257  0.83723164     3  1     1 1.4041769
2  0.45296224  0.25112540     3  1     1 1.3431673
3 -0.25915599 -0.62759966     1  0     3 5.5585029
4  0.02984766 -0.38707173     2  2     3 1.7036150
5 -0.24467357  0.71313728     1  0     3 0.1297077
6 -0.16332535  0.02312365     1  0     3 6.0128203
```

In the next step, a regression tree for the circular response is fitted employing the covariates `x1` and `x2` as potential splitting variables. The results in Figure 1 show that the fitted parameters are very close to the real values of the three respective groups. The total log-likelihood is -1140.111 with 8 degree of freedom.

```
R> m.circtree <- circtree(y ~ x1 + x2, data = sdat)
R> #logLik(m.circtree)
```

4. Summary

Circular response variables are common in a variety of application. However, few regression methods and no distribution trees are so far implemented for circular response values in R.

This paper presented the **circtree** package that provides functions to both circular regression models and to distribution trees employing a von Mises distribution. The main functions are illustrated for artifical data, however, many more exported functions and methods are not shown in this short summary paper and are ready for testing. Additionally, a shiny app is implemented for illustrating the von Mises distribution and exported families for **bamlss** and **disttree** are provided for comparison.

References

Banerjee A, Dhillon IS, Ghosh J, Sra S (2005). "Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions." *Journal of Machine Learning Research*, **6**(September), 1345–1382. URL http://jmlr.csail.mit.edu/papers/v6/banerjee05a.html.

Chambers JM, Hastie TJ (1992). *Statistical Models in S*. Chapman & Hall, London.

Fisher NI, Lee AJ (1992). "Regression Models for an Angular Response." *Biometrics*, **48**(3), 665–677. doi:10.2307/2532334.

Hothorn T, Zeileis A (2015). "partykit: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research*, **16**, 3905–3909. URL http://jmlr.org/papers/v16/hothorn15a.html.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. URL http://www.jstatsoft.org/v34/i01/.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 1345–1382. URL http://jmlr.csail.mit.edu/papers/v6/banerjee05a.html.

Affiliation:

Moritz N. Lang
Universität Innsbruck
6020 Innsbruck, Austria
E-mail: moritz.n.lang@gmail.com